
# Análise de comentários de clientes com o auxílio a técnicas de *Text Mining* para determinar o nível de *(in)*satisfação

por

Ana Catarina Barbosa Forte

**Tese de Mestrado em Modelação, Análise de Dados e Sistemas de Apoio à Decisão**

Orientador: Pavel Bernard Brazdil

Coorientador: João Manuel Portela da Gama

**Faculdade de Economia**

Universidade do Porto

2015

# Análise de comentários de clientes com o auxílio a técnicas de *Text Mining* para determinar o nível de *(in)*satisfação


por

Ana Catarina Barbosa Forte


## Tese de Mestrado em Modelação, Análise de Dados e Sistemas de Apoio à Decisão

Tese submetida à Faculdade de Economia da Universidade do Porto para obtenção de grau de Mestre em Modelação, Análise de Dados e Sistemas de Apoio à Decisão


Orientador: Pavel Bernard Brazdil

Coorientador: João Manuel Portela da Gama

**Faculdade de Economia**

Universidade do Porto


2015

# Nota Bibliográfica

Ana Catarina Barbosa Forte nasceu em Viana do Castelo, no dia 30 de Abril de 1991. Concluiu a sua formação básica e secundária em Viana do Castelo. Terminou a licenciatura em Gestão, na Faculdade de Economia da Universidade do Porto, em Julho de 2013. Em Setembro do mesmo ano, iniciou o Mestrado em Modelação, Análise de Dados e Sistemas de Apoio à Decisão na mesma instituição.

Esta dissertação surgiu em consequência de uma colaboração com uma empresa de telecomunicações.

# Agradecimentos

Quero agradecer ao meu orientador, Pavel Brazdil, que sempre mostrou disponibilidade e boa disposição para acompanhar todo o processo de construção desta dissertação e que me incentivou a exigir mais de mim própria através das suas críticas construtivas.

Agradeço ao meu coorientador, João Gama, pelas sugestões prestadas nas reuniões e por ter sido o responsável por estabelecer esta parceria entre empresas e alunos, fomentando um grande contributo para a minha formação.

Gostaria também de agradecer, à empresa de telecomunicações, por ter tido a oportunidade de trabalhar num projeto aliciante e por me fornecer todos os recursos necessários à elaboração deste documento, bem como por ter sido uma experiência verdadeiramente enriquecedora.

Um profundo obrigado aos meus pais por apostarem sempre na minha educação e pelo apoio incondicional prestado ao longo desta jornada.

Resta-me agradecer a todas as pessoas que passaram pelo meu percurso e que me ajudaram a superar momentos difíceis, um especial obrigado a André Azevedo.

# Resumo

Um problema com que as organizações se deparam é como armazenar grandes quantidades de dados e tirar proveito deles. Devido aos avanços tecnológicos, observamos que estamos expostos a uma infinidade de informações e portanto pesquisar e extrair o mais importante torna-se uma tarefa essencial e cada vez mais difícil. Neste trabalho, o conjunto de dados utilizado é constituído por notas dos assistentes de apoio ao cliente, fornecido por uma empresa de telecomunicações portuguesa. O problema reside no facto de tentar identificar quais os clientes que poderão abandonar a empresa. O objetivo desta dissertação é construir um sistema automático que permite classificar documentos de texto, com base nas opiniões e comentários dos clientes. Inicialmente, procede-se à aplicação de um conjunto de técnicas, na área da Extração de Informação, com o intuito de extrair determinados elementos de texto considerados relevantes nos documentos. Seguidamente, recorre-se à análise de sentimentos para detetar a polaridade das palavras (*negativas, neutras e positivas*) e calculam-se as respetivas pontuações associadas. Os métodos propostos foram devidamente avaliados e os resultados obtidos foram analisados e comparados. Assim, se todos os objetivos tiverem sido alcançados, esta investigação pode abrir novas possibilidades para futuros estudos dentro desta área e também pode mostrar como construir aplicações de interesse para a empresa de telecomunicações e a comunidade em geral.

**Palavras-chave**: Análise de Sentimentos, Classificação Automática, Extração de Conhecimento a Partir do Texto.

## *Abstract*

A problem that organizations face today is that of storing large amounts of data and exploiting it in order to gain some advantage. Due to technological advances, we note that we are exposed to enormous amounts of information and then search and extraction of most relevant items becomes increasingly more difficult. In this work, the data used consists of transcriptions of customer support assistants, provided by a Portuguese Telecommunications Company capturing in essence customer's comments grievances. The problem is that trying to identify which customers may leave the company or if it is possible to prevent this. The purpose of this work is to build an automatic system that permits to classify text documents, based on the opinions and comments of customers. It proceeds by applying a set of techniques in the field of information extraction with aim to extract certain elements considered relevant for further processing. Then, the techniques of sentiment analysis are applied to detect the polarity of the words (*negative, neutral and positive*) and calculate sentiment score of customer's comments. The proposed methods have been properly assessed and the results analysed. Comparisons were also made among different variants of the method with aim to identify the one that achieves the best results. So as all the objectives have been achieved, this research may open new possibilities for future studies in this area and show how to build applications of interest to the telecommunications company and also to the community at large.



**Keywords**: Automatic Classification, Knowledge Extraction from Text, Sentiment Analysis.

# Lista de Abreviaturas

| | |
|---|---|
| CRAN | *Comprehensive R Archive Network* |
| KDT | *Knowledge Discovery from Text* |
| ATM | *Automated Teller Machine* |
| IR | Information Retrieval |
| IE | Information Extraction |
| NLP | Natural Language Processing |
| NER | *Named-Entity Recognition* |
| POS | *Part-of-speech* |
| RSLP | Removedor de Sufixos da Língua Portuguesa |
| KNN | *K-Nearest Neighbours* |
| SVM | *Support Vector Machine* |
| DTM | *Document Term Matrix* |
| BoW | *Bag-of-Words* |
| BoN | *Bag-of-Nouns* |
| TF-IDF | *Term Frequency Inverse Document Frequency* |
| SA | *Sentiment Analysis* |
| pLSA | *Probabilistic Latent Semantic Analysis* |
| LDA | *Latent Dirichlet Allocation* |
| PMI | *Pointwise mutual information* |
| CSNMF | *Constrained Symmetric Nonnegative Matrix Factorization* |

# Índice Geral

# Índice de Figuras

# Índice de Tabelas

# CAPÍTULO 1

# INTRODUÇÃO

O trabalho descrito nesta dissertação pertence à área de *Opinion Mining*, também denominada por *Sentiment Analysis*, que tem como base *Text Mining*.

No seguimento deste trabalho, nomeadamente através destas técnicas, pretende-se chegar a previsões que tenham valor comercial para a empresa em questão.

Neste capítulo começa-se por descrever o contexto e a motivação para esta dissertação. De seguida, descreve-se o problema abordado nesta tese, bem como os seus principais objetivos e recursos utilizados. Por fim, procede-se à organização da estrutura da dissertação.

## 1.1 Enquadramento

Nos últimos anos com a constante inovação tecnológica, como o aparecimento das redes sociais (Facebook, Twitter, Google +, etc.), *blogs* e fóruns, surgiu uma nova "revolução" acerca de como as pessoas se expressam. Estes canais sociais servem de veículo para as pessoas comunicarem as suas opiniões devidamente fundamentadas, como é o caso de pessoas que se sentem profundamente insatisfeitas com algum produto ou serviço, ou mesmo quando estas querem publicar assuntos do foro pessoal. Este período de constante evolução de informação e atualização juntamente com o ingrediente *internet* permitiu que se acumulasse um conjunto massivo de dados textuais. Daí que as empresas estejam cada vez mais preocupadas com o que é dito acerca delas, bem como em extrair conhecimento útil a partir desses dados textuais. Isto exige que as empresas introduzam alterações no seu negócio, sobretudo ao nível do desenvolvimento de uma estratégia diretamente focada em cada cliente, o que nem sempre é fácil.

Por um lado, este impacto da informação teve um lado positivo, pois possibilitou às empresas introduzir e adaptar vantagens competitivas relevantes, pelo facto destas poderem analisar vários fatores como as opiniões dos clientes e os comentários nas redes

sociais, o que permite descobrir padrões que de outra forma passariam despercebidos. Podem-se considerar alguns exemplos desses fatores ignorados como as *notas de apoio ao cliente* ou dos serviços de reparação e as previsões tomando em consideração fatores sazonais. Neste último, atribui-se uma maior atenção aos períodos de férias, no qual se dá especial relevância às previsões de vendas e respetivo crescimento e às campanhas promocionais. Estes agentes influenciadores permitem detetar os produtos e serviços que são mencionados frequentemente em críticas negativas, bem como destacar áreas que deveriam ser otimizadas.

Por outro lado, a informação acarretou uma desvantagem. É de notar que a transformação mais visível que se desenrolou neste contexto, diz respeito ao facto das organizações terem perdido completamente o controlo sobre o que os clientes dizem acerca delas, ou seja, os clientes facilmente publicam as suas opiniões e até comentários negativos e estes são rapidamente divulgados atingindo um grande número de pessoas.

É possível observar que para as empresas existe uma nova visão sobre o cliente, onde é assinalado o seu destaque e relevância. Foram introduzidas boas práticas, tais como ouvir mais e melhor os clientes e aproveitar críticas e elogios para desenvolver novos produtos e por último, o entendimento da importância crescente da manutenção de bons canais de comunicação.

Deste modo, é possível compreender melhor o comportamento dos clientes e tentar prever os seus passos. Um exemplo disso mesmo é verificado quando um cliente deseja comprar um produto ou serviço. Perante esta situação, o cliente consulta as opiniões de outros compradores na internet, por forma a saber até que ponto estes estão ou não satisfeitos com a sua aquisição. Com base nesta informação, decide se finaliza ou não a compra. Este comportamento social por parte de cada cliente tornou-se muito frequente, influenciando e melhorando a sua tomada de decisão.

Desta forma, torna-se particularmente relevante o volume diário de informação sob a forma de texto que as empresas geram, uma vez que necessitam de compreender, tratar e avaliar corretamente essa informação. Assim, as organizações podem ganhar vantagem competitiva se analisarem estes dados.

Com determinadas ferramentas de *Text Mining* as organizações podem, deste modo, selecionar as informações que permitem gerar valor acrescentado e consequentemente tomar decisões mais adequadas. Esta técnica é bastante útil, uma vez que a maior parte da informação de negócio das empresas está sob a forma de texto, maximizando, assim, o valor da própria organização e potenciando uma boa interação entre a empresa e o cliente.

Para uma melhor compreensão acerca do conteúdo dos documentos fornecidos no âmbito desta dissertação, segue-se uma breve explicação do serviço prestado por empresas de telecomunicações.

As empresas costumam ter um serviço de solicitações – *Service Request* – é um serviço que envolve questionários de satisfação, campanhas promocionais, resoluções de problemas, mudanças de palavras passes, para que os clientes habituais e potenciais expressem os seus comentários, opiniões e dúvidas. Este serviço é efetuado pelos assistentes técnicos que estão nos seus postos de trabalho, a atender chamadas telefónicas dos clientes e ao mesmo tempo a escrever os problemas que as pessoas detetam nos produtos ou serviços subscritos. São precisamente esses documentos escritos pelos assistentes de apoio ao cliente que evidenciam aspetos importantes para os clientes e para a empresa, ou ainda registam o grau de satisfação dos mesmos, bem como detetam os sentimentos associados à relação que o cliente possa ter com a empresa. Quase se pode dizer, que não analisar estes documentos de texto é o mesmo que não ouvir a "voz" do cliente!

Para que sejam possíveis as tarefas de classificação e de extração de opiniões torna-se necessário saber quais os termos relevantes que se encontram nos relatórios feitos pelos técnicos, que neste caso, são termos relativos aos equipamentos e serviços da empresa de telecomunicações: telefone, telemóvel, televisão, internet, cinema, pagamentos/faturação e apoio/ajuda ao cliente, entre outros. Note-se que uma boa análise destes documentos só pode ser realizada com base num bom conhecimento desta unidade de negócio, proveniente da relação cliente/empresa.

Os documentos de texto que possuem estas anotações dos técnicos, na maior parte dos casos, não contêm qualquer estrutura implícita. Estes documentos são escritos em

linguagem natural, onde a informação apresentada é não estruturada. Embora para os clientes seja mais fácil e prático deixar este tipo de informação sob a forma de texto livre, é mais difícil para as empresas processá-la com *software* com o objetivo de extrair alguma informação útil.

Em conclusão, nas empresas de telecomunicações é útil perceber o potencial destes dados, pois pode ter impacto em diversas áreas, como por exemplo em adquirir informações sobre a fidelização e lealdade do cliente, em efetuar previsões com o intuito de estudar a popularidade ou a introdução de um novo produto ou serviço e em compreender o perfil dos clientes que exprimem opiniões sobre a empresa.

Com estas informações, podemos implementar estratégias personalizadas para cada tipo de cliente, ou seja, reconhecer quais as necessidades do cliente e suas insatisfações com o serviço ou produto, direcionar campanhas exclusivamente para esse cliente e ao mesmo tempo prevenir a perda do mesmo. Deste modo, tem todo o interesse estudar a forma como os clientes se manifestam ao longo do tempo para detetar alterações das suas preferências como, por exemplo, identificar padrões de consumo e possível abandono por parte dos clientes.

Assim as organizações devem procurar estar continuamente atualizadas e com as mais variadas tecnologias, por forma a perceber as necessidades dos seus clientes e a sobreviverem, neste contexto, de competitividade feroz.

## 1.2 Propósito

O texto é o repositório mais importante do conhecimento humano (Lin e Pantel, 2001). Isto significa, que se deve dar uma especial atenção a este tipo de dados, por forma a compreender o que é relevante para o estudo.

O problema abordado nesta tese é a identificação de sentimentos presentes nos documentos de textos. O problema adicional que foi tratado só parcialmente e que podia ser desenvolvido melhor no futuro é a identificação do tópico ou assunto relevante nas frases escritas pelos técnicos.

Os resultados desta análise podem auxiliar na descoberta de produtos ou serviços que precisam de sofrer alterações mais significativas, ou na identificação dos clientes que

estão em risco de abandonar a empresa (*churn*) e no reconhecimento dos clientes que tenham sérias intenções de efetuar uma compra.

O desafio proposto entra diretamente em linha de conta com as notas dos assistentes de apoio ao cliente, isto porque estes textos não foram submetidos a qualquer controlo editorial. Estes dados contêm erros ortográficos e gramaticais não intencionais, abreviações, *emoticons*, palavras e pontuações repetidas para destacar o assunto, exigindo um conjunto de tarefas de pré-processamento extensivas.

Neste trabalho foi realizado um estudo sobre uma empresa de telecomunicações, em que os dados utilizados são reais.

## 1.3 Organização da dissertação

Esta dissertação encontra-se estruturada por cinco capítulos.

O primeiro capítulo é composto pela introdução. O segundo capítulo diz respeito à revisão da literatura, onde são abordados métodos e técnicas de *Text Mining*, seguindo para a exploração da análise de sentimentos, identificação de tópicos, medidas de avaliação e demonstração de alguns exemplos práticos relacionados com esta área de negócio. No Capítulo 3 é realizada uma descrição dos dados em estudo e dos recursos utilizados, assim como se apresenta a metodologia com sistema básico, onde são apresentados e discutidos os seus resultados. Mais tarde, no Capítulo 4 aborda-se a metodologia com sistema avançado, a qual inclui todas as experiências testadas nesta dissertação e que auxiliaram na resolução deste problema. Ainda neste capítulo são avaliados e interpretados os resultados dos objetivos propostos. Por fim, no Capítulo 5 são apresentadas as conclusões e sugestões para trabalhos futuros.

# CAPÍTULO 2

# MÉTODOS DE *TEXT MINING* E DE ANÁLISE DE SENTIMENTOS

Neste capítulo aborda-se a temática do *Text Mining* que é considerada uma área bastante promissora. Este tema é relevante, pois é a base para uma análise de texto, que é um dos objetos deste estudo. *Text Mining* é uma técnica da Inteligência Artificial, que consiste na extração de informação a partir de texto. Basicamente, a extração de texto consiste em transformar palavras ou frases não estruturadas numa forma adequada para poder aplicar técnicas de *Data Mining*.

Os dados não estruturados representam um desafio, pois os documentos de texto muitas vezes são inconsistentes e ambíguos dado que contém gíria, linguagem específica de determinadas indústrias ou de grupos etários, que podem ser suscetíveis a diversas interpretações. O trabalho desenvolvido, nesta dissertação, enquadra-se neste contexto.

## 2.1 *Text Mining*

Existe ouro escondido nos dados de cada empresa – a extração de conhecimento a partir de texto promete ajudar as organizações a encontrá-lo (Dörre et al., 1999). É de realçar que apenas uma minoria das informações das organizações é que se encontra estruturada, pois habitualmente estas informações estão disponíveis nas bases de dados da empresa. No entanto, existem imensos dados não estruturados que não estão a ter a devida atenção ou que simplesmente passam despercebidos. Tem-se como exemplo as cartas dos clientes, os contractos, o correio eletrónico, as gravações de chamadas telefónicas, os registos de reparações, as dúvidas sobre produtos, os registos dos empregados, entre outros. Uma solução que permite retirar o ouro escondido dos dados, passa precisamente pela extração de conhecimento e de informações a partir do texto.

*Text Mining*, ou KDT *(Knowledge Discovery from Text*) pode ser definido como um processo de conhecimento intensivo, em que um utilizador interage com uma coleção de

documentos ao longo do tempo, por meio de um conjunto de ferramentas de análise (Feldman e Sanger, 2006).

A internet é um grande incentivo a esta temática, uma vez que existem grandes volumes de texto nas redes sociais (*Facebook, Instagram, Twitter),* no *email*, nos *blogs*, tal como nos motores de busca como a Google, Yahoo e Bing. É necessário analisar o conteúdo destes dados e retirar as informações que adicionam mais valor.

Esta tarefa não é fácil, pois o texto contém muito "lixo". Além disso, as pessoas escreverem com erros ortográficos, palavras que se apresentam na forma soletrada ou juntas, abreviações, ícones de emoção, símbolos, palavras longas para enfatizar a situação e pontuação aleatória ou repetida. É normal que um texto contenha também muitos sinónimos ou mesmo palavras com diferentes significados. A título de exemplo, a palavra "*caixa*" pode desempenhar diversos significados como: *ATM (Automated Teller Machine)* ou a caixa de multibanco, caixa para guardar ou transportar alguma coisa, compartimento para proteger um mecanismo como a caixa do elevador, registo de pagamentos e recebimentos, ou cofre para guardar dinheiro, entre outros.

### 2.1.1 O que é um documento?

A definição de documento é subjetiva. O documento pode ser constituído apenas por uma frase, ou por outro lado ser um conjunto de várias páginas. Uma coleção de documentos é denominada de *corpus*.

### 2.1.2 Técnicas de *Text Mining*

A análise de texto envolve um conjunto de processos visando obter informações importantes de um texto. Seguidamente, algumas tarefas típicas de *Text Mining* são descritas.

**Procura/Seleção de Informação (*Information Retrieval*)**: Dentro de um possível conjunto de documentos (universo), pretende-se unicamente a extração dos documentos mais importantes. O que realmente importa é o tempo, por isso, é de todo conveniente encontrar informação de forma automática no menor tempo possível.

**Extração de Informação (*Information Extraction*)**: Ao contrário do tópico anterior, o que interessa aqui são os elementos importantes que estão contidos em determinados documentos. Analisa-se o texto sem restrições, em que o objetivo é extrair dos documentos determinados factos específicos, entidades ou relacionamentos.

**Processamento de Linguagem Natural (*Natural Language Processing, NLP*)**: Tem como objetivo analisar a linguagem escrita e falada, com o intuito de convertê-la para uma representação mais formal, e assim torná-la manipulável por programas de computador (Nunes et al., 2013).

### 2.1.3 Estruturas possíveis de documentos

A extração de informação de texto consiste em extrair a informação a partir de dados semiestruturados ou não estruturados. A figura 2.1 mostra um exemplo de dados estruturados, semiestruturados e não estruturados.


**Figura 2.1** – Exemplos da estrutura de dados (adaptado de Mike Bergman, 2005).

Os dados não estruturados não assumem qualquer tipo de formatação, pois normalmente estão na forma de texto livre. O texto possui estrutura, mas essa é apenas uma estrutura que segue regras linguísticas destinadas para os humanos e não para os computadores. Como foi referido anteriormente, muitos documentos inserem-se nesta categoria.

Os semiestruturados apresentam alguma estrutura como, por exemplo, o modelo europeu de apresentação do *Curriculum Vitae*.

Quanto aos dados estruturados, estes já obedecem a um conjunto de regras e restrições rígidas. As bases de dados relacionais são um dos possíveis exemplos deste tipo de dados.

### 2.1.4 Pré-processamento

Com a aplicação de técnicas de pré-processamento são verificadas alterações no documento inicial, que podem provocar uma melhoria na qualidade dos resultados obtidos. Seguidamente, serão detalhadas algumas tarefas de pré-processamento efetuadas neste trabalho.

**Segmentação de texto (*Tokenização*)**: *Token* é o nome que se dá aos termos extraídos dos textos, sejam eles palavras ou expressões compostas. Um *token* pode ser um conjunto de n (n=1,2,… n) caracteres consecutivos, ao qual se denomina *n-gram*. No entanto, a abordagem mais usual é que os *tokens* extraídos do texto sejam palavras, nesse caso, as palavras são separadas por espaços ou sinais de pontuação que são considerados delimitadores de *tokens*. No caso de o objetivo ser extrair frases, a pontuação (ponto final, exclamação, interrogação, vírgula, entre outros) será a fronteira ou *token* final da frase.

O objetivo da segmentação de texto, tanto ao nível da palavra como da frase, é transformar o texto num conjunto de *tokens*.

**Remoção de números**: Eliminam-se estes termos pelo facto de não acrescentarem informação relevante a muitas tarefas de *Text Mining*.

**Transformação de letras maiúsculas em minúsculas**: Este processo permite identificar palavras iguais, mas escritas de forma diferente, como por exemplo "*Telecomunicações*", "*TELECOMUNICAÇÕES*" e "*telecomunicações*". O resultado final será "*telecomunicações*" em todos os casos.

**Remoção de *stopwords***: *Stopwords* são palavras e frases comuns que não acrescentam qualquer informação a o texto, sendo as mais frequentes preposições, artigos, e pronomes, como por exemplo "um", "uma", "e" que são palavras irrelevantes (Sedbrook e Lightfoot, 2010).

Uma das vantagens acarretadas pela remoção destas palavras comuns é a redução de processamento devido à considerável diminuição do tamanho do vetor de termos.

***Stemming***: Este processo visa a obtenção da raiz morfológica, eliminando os prefixos e sufixos que carregam a informação gramatical ou lexical da palavra (Moral, C. et al.,

2014). Este passo tem como objetivo reduzir palavras que se encontram em formas derivadas, para a sua forma base (Gomes, H., 2012). Exemplo disso é a transformação das formas "*enterrar*", "*terramoto*" e "*aterrar*" para a sua forma base "*terra*". O estudo de Hull comprovou que a utilização dos algoritmos de redução morfológica pode gerar uma melhoria no desempenho do processamento em aproximadamente 5% (Hull e Grefenstette, 1996).

Notemos que um dos efeitos decorrente do *stemming* é reduzir o número de termos distintos num *corpus* de texto e aumentar a frequência da ocorrência da raiz dos termos. Para algoritmos de classificação que têm em conta a frequência, às vezes isto pode fazer a diferença (Weiss et al., 2010).

Um problema dos algoritmos de *stemming,* que utilizam remoção de prefixos/sufixos, é que estes são dependentes das respetivas línguas, pois baseiam-se diretamente nas regras de formação de palavras de cada língua (Honrado et al., 2000).

Os algoritmos que se destacam na língua portuguesa são o algoritmo de *Porter* (2005) na versão portuguesa, o "*Removedor de Sufixos da Língua Portuguesa*" (RSLP), proposto por Orengo e Huyck (2001) e o algoritmo *STEMBR*, proposto por Alvares, Garcia e Ferraz (2005).

Depois das tarefas anteriores, ainda se pode adicionar um conjunto de tarefas de maior grau de complexidade, que podem melhorar o desempenho final dos algoritmos. A seguir, apresentam-se algumas dessas tarefas avançadas.

**Remoção de marcações de HTML**: Se os documentos estiverem em HTML, o objetivo é remover as marcações, normalizando os documentos e eliminando as formatações de texto que se encontrem a negrito, sublinhado ou em itálico, tamanhos de letra, espaços em branco, pontuação, símbolos (caracteres não alfanuméricos), entre outros atributos que possam ser descartados.

**Lidar com abreviações**: Uma vez que os assistentes de apoio ao cliente falam e escrevem ao mesmo tempo com os clientes, muitos dos textos aparecem com abreviaturas. Uma solução consiste em adicioná-los ao léxico criado, convertendo esses termos para a sua forma correta.

**Redução de caracteres repetidos**: Normalmente as pessoas utilizam letras repetidas para intensificar o sentimento, como "adooooooooro o pacote de canais". Neste caso, devem-se adotar regras, como por exemplo dizer que os caracteres que estejam repetidos mais do que duas vezes são eliminados, ou então, simplesmente reduzir-se a palavra à sua forma original. No caso, do exemplo acima passaria para "adooro o pacote de canais", esta palavra "adooro" deveria ser acrescentada ao nosso léxico.

**Substituir ícones de emoção**: Este passo consiste na substituição destes ícones pelo seu nome, como por exemplo ":)" corresponde à emoção "contente" ou ainda ":s" significa "triste".

**Identificação de termos compostos**: A fase de identificação de termos compostos denomina-se por *Multi-Word Units* (Altenberg, 1998) ou *fraseologia* (Cowie et al., 1998). Tem-se como exemplo as palavras "*guarda-chuva*" e "*Engenharia Gestão Industrial*" que se forem tratadas simplesmente como um conjunto de palavras separadas, ou seja, "*guarda*" e "*chuva*" ou "*Engenharia*", "*Gestão*" e "*Industrial*", o sentido inicial dessas palavras é modificado. Uma das formas de identificação destes termos é feita com base na frequência em que ocorrem numa dada coleção de documentos. Outra hipótese é extrair os termos compostos com recurso a um dicionário de expressões.

**Análise sintática (*Part-of-speech* - POS)**: Note-se que a mesma palavra pode ter classes gramaticais diferentes, ou seja, essa palavra pode ser um *substantivo*, *verbo* ou *adjetivo*, dependendo da posição que ocupa na frase. A classe gramatical pode ser tomada em consideração na atribuição de valor de sentimento.

Os principais algoritmos de análise sintática de hoje em dia atingem níveis de acerto superiores a 97% (Manning, 2011).

Um dos trabalhos mais recentes em Portugal é o SentiLex-PT02, que é um léxico de palavas de sentimento portuguesas que contém lemas, classes gramaticais, atributos de sentimento e observações específicas do anotador, desenvolvido por Silva, Carvalho e Sarmento (2012). Este léxico é disponibilizado mediante pedido.

**Reconhecimento de entidades nomeadas (*Named Entity Recognition*)**: Uma entidade pode ser um produto, serviço, pessoa, locais, organização ou evento, no documento de opinião (Liu, 2012).

Identificar as entidades para uma empresa pode ser valioso, visto que se pode identificar os concorrentes de mercado. A procura de palavras em que a primeira letra é maiúscula auxilia o reconhecimento do nome das entidades como, por exemplo, "*Obama*", "*TAP*", entre outras.

### 2.1.5 Representação de Texto

Existem classificadores, como as árvores de decisão, KNN, *Naive Bayes*, SVM, entre outros, requerem que os dados sejam apresentados em forma de tabelas. Nesta representação cada linha contém informação sobre um documento e cada coluna representa um atributo em específico (um termo). Estas tabelas são normalmente referidas como *Document Term Matrix* (DTM).

Resultados experimentais mostraram que representações de texto mais sofisticadas perdem desempenho em relação à abordagem *bag-of-words* (Apté, et al.,1994).

A representação *Bag of Words* (BoW) consiste em transformar o texto em palavras que ocupam certa posição no vetor. Cada elemento do vetor representa um termo (Amine et al., 2010).

A representação pode ser binária, que simplesmente determina se o dado termo ocorreu no documento. Uma outra representação usa a frequência da palavra no documento.

A representação TF-IDF (*Term Frequency-Inverse Document Frequency*) considera a frequência e peso (ponderação) de cada palavra. Em (2.2) apresentamos a fórmula que permite calcular $w_{ij}$,

$$w_{ij} = f_{ij} * log\left(\frac{N}{n_i}\right) \tag{2.1}$$

em que $f_{ij}$ representa a frequência da palavra $i$ no documento $j$, $N$ é o número total de documentos e $n_i$ mostra o número de documentos que possuem a palavra $i$ .

### 2.1.6 Representação de texto: *Bag-Of-Nouns*

A maioria dos modelos de representação usam palavras-chave, como *Bag-Of-Words* (BoW), onde os documentos são reproduzidos como um conjunto de termos desordenados. Por exemplo, as frases "*o urso come homem* '" e "*homem come urso*" são idênticas, contudo estas apresentam significados diferentes.

Uma melhor representação dos dados é a de *Bag-Of-Nouns* (BoN) que considera apenas os substantivos, incluindo nomes próprios e objetos.

A representação *Bag-Of-Nouns* trata cada exemplo, como um vetor ***não ordenado***, composto por todos os substantivos. Os substantivos são excluídos, caso sejam encontrados na lista de *stopwords*. Os substantivos não passam pela tarefa de pré-processamento, *stemming* (como na representação *Bag-Of-Words*), visto que não é favorável para esta análise.

Existem dois motivos muito fortes para se enveredar por esta representação.

Em primeiro lugar, os substantivos são considerados mais valiosos do que outras categorias de palavras, pois indicam o que é tratado no documento. No domínio financeiro, outras categorias podem ser importantes, como por exemplo os verbos "*vender*" e "*comprar*", mas no domínio da opinião dos consumidores tem mais ênfase as pessoas, os objetos e os eventos.

O segundo motivo deriva do facto dos substantivos serem apenas uma pequena fração de palavras numa frase, pois o espaço de características de *Bag-Of-Nouns* seria apenas uma fração do espaço de características de *Bag-Of-Words*.

### 2.1.7 Identificação de Termos Informativos

Após a transformação de documentos em BoW, o conjunto de termos pode ser bastante numeroso. Assim, pode-se estar interessado em eliminar certas palavras que não têm importância nesta análise. Com a seleção de atributos as palavras irrelevantes são removidas.

Assim, não precisamos de tantos recursos computacionais, pois a dimensão dos textos é reduzida. Por outras palavras, as características existentes são transformados num espaço

dimensional inferior e, assim, a compreensibilidade pode ser significativamente melhorada (Feldman e Sanger, 2007).

A este propósito, existem medidas úteis que têm em consideração as relações entre características e categorias, como por exemplo *pointwise mutual information* (PMI) e *information gain*, entre outros, que são a base de algumas técnicas de eliminação de atributos.

O critério PMI identifica pares de palavras que tendem a aparecer juntas. Logo, estas palavras apresentam uma relação semântica (Oakes, 2014).

$$PMI(Termo_1, Termo_2) = log_2\left(\frac{P(Termo_1 \wedge Termo_2)}{P(Termo_1)\,P(Termo_2)}\right) \quad (2.2)$$

Este rácio $\frac{P(Termo_1 \wedge Termo_2)}{P(Termo_1)\,P(Termo_2)}$ representa o grau de dependência estatística entre os dois termos. O cálculo de PMI corresponde à quantidade de informação que se adquire sobre a presença de uma das palavras, quando se observa a outra.

O critério do ganho de informação (*Information Gain, IG*) mede o número de bits de informação obtida para a predição de categorias, pelo conhecimento da presença ou ausência do termo num documento (Yang e Pedersen, 1997). Só os atributos com maior valor de informação é que são preservados.

Tendo em conta duas distribuições de probabilidade X e Z, diz-se que esta medida representa a quantidade de acréscimo de bits necessários a Z para que se consiga utilizar X na construção das suas mensagens. Para determinar esta medida é necessário calcular a entropia de uma distribuição de probabilidade. A entropia pode ser calculada em (2.3) da seguinte forma,

$$H(X) = -\sum_{i=1}^{n} P(X)_i * log_2\ P(X)_i \quad (2.3)$$

em que $H(X)$ é a entropia da distribuição de probabilidade de $X$, $P(X)$ representa a probabilidade de acontecimento de cada valor de $X$ e $n$ é o número de valores que a variável $X$ pode tomar. A segunda distribuição é uma particularização da primeira, ou

seja, é a primeira distribuição condicionada a um dos seus valores, neste caso está representado por $Y$, em (2.4).

$$H(Z) = -\sum_{i=1}^{n} P(Y)_i \sum_{j=1}^{n} P(X|Y)_j * \ log_2 \ P(X|Y)_j \tag{2.4}$$

O ganho de informação é considerado como a diferença entre entropias das duas distribuições representado em (2.5).

$$IG(X|Y) = H(X) - H(Z) =$$

$$-\sum_{i=1}^{n} P(X)_i * \ log_2 \ P(X)_i \ + \sum_{i=1}^{n} P(Y)_i \sum_{j=1}^{n} P(X|Y)_j * \ log_2 \ P(X|Y)_j \tag{2.5}$$

Esta medida corresponde ao ganho de informação que $Y$ possui para a determinação de $X$.

## 2.2 Análise de Sentimentos

Análise de sentimentos, também é chamada de *opinion mining*, é um campo de estudo que analisa as opiniões das pessoas, sentimentos, avaliações, atitudes e emoções em relação a entidades como produtos, serviços, organizações, pessoas, problemas, eventos, tópicos e atributos (Liu, 2012).

As informações contidas no documento de texto podem ser objetivas, subjetivas ou ambas. Normalmente no texto subjetivo encontram-se opiniões positivas e negativas, enquanto no texto objetivo são descritos factos. Uma frase objetiva apresenta algumas informações factuais sobre o mundo, enquanto uma frase subjetiva expressa alguns sentimentos pessoais, opiniões ou crenças (Liu, 2012).

Com a análise de sentimentos identifica-se a subjetividade e objetividade presente no texto e a polaridade que está associada ao texto subjetivo. A polaridade ou orientação de um texto pode ser positiva, negativa ou uma mistura destes. Considera-se uma polaridade neutra quando o texto é objetivo.

### 2.2.1 Atribuição de sentimento a documentos

Segundo o autor Bing Liu (2006), existem três avaliações fundamentais para documentos de texto:

**Classificação do sentimento (*Sentiment Classification*):** O problema é considerado como um problema de classificação. No entanto apesar da classificação de texto e de sentimento serem similares, existe uma ligeira diferença. Na classificação de texto procura-se atribuir a um documento, uma ou mais classes, de acordo com determinadas características ou atributos. Procede-se de forma análoga na classificação de sentimentos, mas procuram-se expressões que capturem a opinião do autor ou das pessoas envolvidas.

Neste trabalho tratamos, essencialmente, uma classificação efetuada ao nível do documento. Neste nível pretende-se determinar se um documento de opinião expressa um sentimento positivo ou negativo (Pang et al., 2002).

Quando se expressa uma opinião sobre um determinado produto, serviço, individuo ou qualquer outra entidade pode-se identificar a polaridade associada, como por exemplo "a qualidade de atendimento do funcionário era **terrível**" (*sentimento negativo*) ou "a capa de proteção do telemóvel é **muito resistente**" (*sentimento positivo*).

Como se pode notar este tipo de classificação é um pouco limitado. Apenas classifica se um documento é positivo ou negativo, ou seja, não consegue identificar se as pessoas gostaram ou não das características.

**Atribuição de sentimento a entidades ou aspetos:** Este processo permite uma análise mais profunda que o processo anterior, uma vez que atua ao nível da frase.

Esta abordagem baseia-se em duas importantes características, ou seja, no reconhecimento dos alvos de opinião (aspetos/objetos) e na identificação dos sentimentos (negativo, neutro ou positivo). Na frase "*a qualidade de atendimento do funcionário é terrível*" pode-se identificar que o foco é a "*qualidade de atendimento"* e o sentimento atribuído a esta característica é *negativo*.

Um texto que é avaliado como positivo sobre um determinado objeto, não significa que o autor tenha opiniões positivas (negativas) sobre todos os aspetos do objeto.

Quando o autor descreve a sua opinião normalmente descreve diversos aspetos (positivos e negativos), aos quais se podem atribuir valores distintos. A opinião global acerca do objeto pode ser numa direção só, positiva ou negativa.

**Frases comparativas e extração de relações (*Comparative sentence and relation mining*)**: As frases comparativas expressam uma relação com base nas semelhanças ou diferenças entre dois ou mais objetos. Estas frases podem assumir duas formas. Utiliza-se a forma *comparativa* para expressar que um objeto apresenta inferioridade ou superioridade em relação a outro, ou a forma *superlativa* para afirmar que o objeto é o melhor ou o pior do que os outros.

### 2.2.2 Classificação do sentimento

Dado um conjunto de documentos $D$, um classificador de sentimentos classifica cada documento $d \in D$ em $N$ classes, positiva e negativa (Saraswat e Patel, 2014).

É comum atribuir uma classe *positiva*, *neutra* ou *negativa* ao texto. Uma aplicação prática desta classificação é descobrir a polaridade ou orientação do sentimento nos textos de acordo com a opinião dos autores, como em comentários de filmes ou descrições de produtos.

### 2.2.3 Abordagens baseadas em léxico

Existem duas abordagens principais para determinar a orientação da opinião sobre cada aspeto ou objeto: a abordagem de aprendizagem supervisionada (*Supervised Learning Approach*) e a abordagem baseada em léxico (*Lexicon-based Approach*). A figura seguinte mostra as diversas alternativas que podem ser seguidas.

Há três maneiras para compilar as listas de palavras de sentimento: a abordagem manual, a abordagem baseada em dicionário e a abordagem baseada numa coleção de documentos.

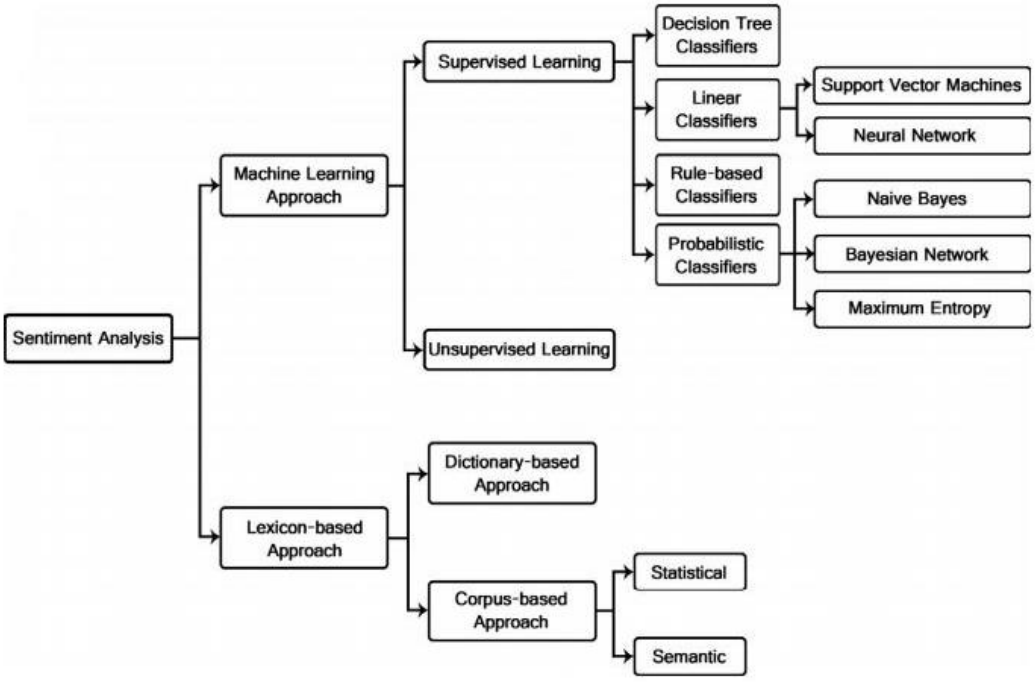



**Figura 2.2** – Técnicas de classificação usadas na análise de sentimentos (reproduzido de Walaa Medhat et al., 2014).

### Abordagem Manual *(Manual Approach)*

A abordagem manual é muito exaustiva, demorada, pois tem que se introduzir todo o vocabulário manualmente. Normalmente este método costuma ser combinado com métodos automáticos para verificação final.

### Abordagem baseada em dicionário *(Dictionary-based Approach)*

No *Dictionary-based approach*, o processo começa com um pequeno conjunto de palavras de opinião que são recolhidas manualmente com as orientações conhecidas. Muitos dicionários contém uma lista de sinónimos e antónimos para cada palavra, como por exemplo *WordNet* (Miller et al. 1990) ou *Thesaurus*. O método pesquisa sinónimos ou antónimos e se encontrar novas palavras vai adicioná-las à lista inicial de palavras. O processo termina assim que não são encontradas novas palavras.

Ghahramani e Zhu (2002) usaram o algoritmo de propagação em classes (*Label Propagation Algorithm*) com algumas alterações, para avaliar cada palavra. Primeiro, a cada semente positiva é dada a pontuação +1, a cada semente negativa é atribuída a pontuação de -1 e às restantes palavras são dadas pontuações de 0. As pontuações são verificadas durante o processo de propagação. Depois de um dado número de iterações,

a propagação termina. Com base na escala logarítmica, calculam-se as pontuações finais consoante os seus graus (positivo ou negativo).

Turney e Littman (2003) empregaram o critério *Pointwise mutual information* (PMI) para calcular a orientação semântica de uma palavra. Este método calcula a orientação da palavra com base na força de associação a um conjunto de palavras positivas menos a força de associação a um conjunto de palavras negativas. A força de associação é medida utilizando o PMI.

Park e Peng (2011) criaram um léxico de sentimento usando *constrained symmetric nonnegative matrix factorization* (CSNMF). Os autores usaram *bootstrapping* para encontrar um conjunto candidato de palavras de sentimento no dicionário e de seguida calcularam a orientação semântica de cada palavra, com base numa coleção grande de documentos.

O *Dictionary-based approach* apresenta uma grande desvantagem que é a incapacidade de identificar palavras de opinião que diferem a sua orientação em determinados contextos. Por exemplo, a orientação da palavra "velho" depende do contexto. Se é aplicada à qualidade do vinho, habitualmente é considerada como positiva. No entanto, se estivermos a referir-nos a um idoso pode ser considerada como uma expressão negativa.

**Custos com as abordagens baseadas em léxico**

Com um léxico completo, o custo para a análise de texto automatizada é extremamente baixo (Quinn et al., 2010).

Os léxicos já existentes oferecem facilidade de uso e têm sido aplicados em diversos contextos, contudo como são frequentemente dependentes do contexto, levam potencialmente a sérios erros de investigação (Grimmer and Stewart, 2013).

A construção de diferentes léxicos para cada análise é possível, mas os custos de construção de um dicionário são sempre elevados (Gerner et al., 1994; Quinn et al., 2010), bem como validar o dicionário pode ser difícil (Grimmer and Stewart, 2013).

### 2.2.4 Análise dos sentimentos baseada em aspetos (*Aspect-Based SA*)

Por vezes, a classificação ao nível do documento ou da frase não é suficiente, uma vez que não identifica os alvos de opinião ou os sentimentos sobre esses alvos. O alvo de opinião (*opinion target*) é decomposto em *entidades* e *aspetos*.

Um documento em geral pode ser classificado como positivo, quando se refere a uma determinada entidade. No entanto, isso não significa que o autor tenha opinião positiva sobre todos os aspetos dessa entidade. O mesmo se passa com um documento de opinião negativa, isto é, não quer dizer que o autor ache que tudo é negativo.

Portanto, uma análise completa pretende determinar os aspetos e descobrir a orientação do sentimento em relação a cada objeto.

Existe um conjunto de seis etapas para descobrir detalhadamente estes componentes. Contudo, apenas duas etapas merecem destaque neste estudo ao nível do aspeto.

A primeira etapa compreende a extração dos aspetos que foram avaliados, como por exemplo "*a qualidade das aplicações do telemóvel é excelente*", em que o aspeto é a *"qualidade das aplicações*", sendo o objeto representado pelo "*telemóvel*".

Na segunda etapa, pretende-se determinar se a opinião sobre os vários aspetos é *negativa*, *neutra* ou *positiva*. No exemplo anterior, vê-se que o aspeto "*qualidade das aplicações*" está associado a um sentimento *positivo*.

Portanto, caso os alvos de opinião (entidades e aspetos) sejam fornecidos não é necessário proceder à etapa de extração dos aspetos, apenas seria preciso identificar os sentimentos dos alvos de opinião.

Os adjetivos desempenham um papel fundamental, uma vez que estes descrevem características específicas ou propriedades (atributos) da entidade. A título de exemplo tem-se a frase "*este telemóvel é barato*", em que se pode constatar que a palavra "*barato*" representa a orientação do sentimento e também indica que o aspeto é o "*preço*", assim como o adjetivo "*magro*" descreve o aspeto "*peso*". No entanto, as expressões podem ser bem mais complexas, como por exemplo "*o telemóvel não vai caber facilmente no bolso*", em que "*no bolso*" revela o aspeto "*tamanho*".

### 2.2.5 Lidar com a negação

A negação é um recurso linguístico muito comum que afeta a polaridade e, portanto, precisa ser levada em consideração nesta análise. A negação não é transmitida apenas por palavras de negação comuns como por exemplo "*não*" ou "*nem*", é também afetada por outros recursos lexicais.

Algumas pesquisas mostraram que existem muitas palavras que podem inverter a polaridade de uma opinião, como *Valence Shifters* e conectores das frases (Polanyi et al., 2006).

Na negação é necessário identificar o seu âmbito, isto é, a negação pode ser local (*por exemplo, **não** é bom*), ou pode envolver dependências de distância longa (*por exemplo, **não parece** muito bom*) ou a negação do sujeito (*por exemplo, **ninguém** pensa que é bom*). Por vezes, em vez de negar, até pode intensificar como por exemplo ***não só** é bom, **mas também** é incrível* (Wilson et al., 2005). Para descobrir o objetivo da negação, a sequência de palavras na frase deve ser identificada.

As expressões de negação nas frases podem ser feitas com recurso a verbos, advérbios, sufixos ou prefixos. Estas expressões podem ocorrer mais do que uma vez na frase e também podem cancelar uma expressão e dar um significado negativo, a título de exemplo "*Eu **não** posso ficar **sem** satisfação*" (Jagger e Richards, 1965).

Os *valence shifters* são expressões que alteram a polaridade do sentimento. Palavras que advém da negação como "*não*", "*nunca*", "*ninguém*", "*em nenhum lugar*", "*nenhum dos dois*" e "*não pode*", são os tipos mais comuns de *valence shifters*. Também nos conectores adversativos como "*mas*", "*no entanto*", "*todavia*", "*porém*", "*contudo*", "*entretanto*", "*senão*", "*não obstante*" e expressões como "*exceto*" ou "*exceto para*", verifica-se uma ideia de contraste ou oposição que deve ser levada em consideração.

Os verbos auxiliares modais, como "*ser*", "*poder*", "*obrigar*", "*dever*", "*acabar*", "*terminar*", "*ter de*", "*crer*", "*precisar*", "*saber*" e adjetivos como "*capaz*", "*talvez*", "*possível*", "*provável*", "*necessariamente*", são outras expressões que podem alterar a orientação do sentimento.

No entanto, é preciso ter cuidado porque nem sempre acontece esta alteração. Advérbios como "*mal*" e "*dificilmente*", palavras como "*falhar*", "*omitir*", "*negligência*" e o sarcasmo ou ironia também alteram o sentido das palavras de opinião.

No entanto é preciso ter cuidado, porque existem outas expressões em que esta mudança não se verifica, como os conectores copulativos que estabelecem uma ideia de adição ou acréscimo, como por exemplo "*não*" em "*não só… mas também*", "*e*", "*nem*", "*mas também*", "*ainda*", "*como*", etc.

Posto isto, a análise de negação pode ser realizada com recurso a *part-of-speech*, *Bag-of-Words*, *Dependency Tree,* entre outros. No entanto, estes recursos quando combinados proporcionam melhores resultados.

### 2.2.6 Deteção de sarcasmo

O sarcasmo é uma forma sofisticada de dizer ou escrever o oposto do que realmente se quer dizer, ou seja, é uma ironia. Pode-se dizer que a ironia é quase como uma inversão do significado real num determinado contexto. Normalmente, é muito difícil de lidar com este género de expressões.

González-Ibáñez et al. (2011) estudaram o problema no contexto da análise de sentimento, utilizando dados do *Twitter*. O objetivo, deste estudo, era diferenciar os *tweets* sarcásticos dos *tweets* não sarcásticos, em que não foram consideradas as expressões neutras. Recorreram ao uso de uma abordagem de aprendizagem supervisionada através de SVM e regressão logística. As características utilizadas foram à base de *unigrams* e de informação baseada em dicionário. Neste último, recorreram a dicionários de categorias de palavras (Pennebaker), *Wordnet Affect* e uma lista de interjeições e de pontuação.

Os resultados obtidos por três classes (sarcástico, positivo e negativo) revelaram que o problema é bastante complexo, uma vez que a precisão foi apenas de 57%.

### 2.2.7 Agrupar significados em categorias similares

As palavas ou frases que evidenciam significados semelhantes, precisam de ser agrupadas em categorias de sinónimos como, por exemplo, "*áudio*" e "*altifalante*" que se referem à mesma característica "*som*". Para a análise de sentimentos identificar e agrupar estas

características que apresentam significados semelhantes facilita a visualização e compreensibilidade dos dados em estudo.

| Aspect | Synonyms |
|---|---|
| battery | battery life;talk time; |
| size | screen size;inch;width;height;diagonal; |
| memory | RAM;gb;mb;storage;hard disk; |
| display | screen;graphics;color; |
| resolution | pixel;camera resolution; |
| camera | picture;photo; |
| weight | |
| sound | audio;speaker;speakerphone; |
| music | mp3;mp3 player; |
| video | movie;film; |
| earphone | headphone; |
| keyboard | querty;button;keypad;typing; |

**Figura 2.3** – Lista de objetos com os seus sinónimos criada para telemóveis e *tablets* (reproduzido de Goga e Stan, 2012).

Também se deve ter em conta que as expressões que descrevem a mesma propriedade podem não ser sinónimas, como por exemplo "*alto*" e "*baixo*" referem-se à mesma característica "*altura*", no entanto, não são sinónimos.

## 2.3 Extração de tópicos

Um tópico ou palavra-chave de um documento é qualquer palavra ou uma sequência de duas ou mais palavras, que resume uma parte do conteúdo desse documento. Por outras palavras, as palavras-chave descrevem os temas principais expressos no documento de texto. A extração de tópicos é bastante útil para documentos de sumarização, classificação e agrupamentos, uma vez que permite um acesso fácil e eficaz à informação relevante.

Os métodos para extração de tópicos podem ser divididos em duas categorias: métodos supervisionados e não supervisionados.

### 2.3.1 Métodos não supervisionados

Os métodos estatísticos de extração de tópicos surgiram como principais métodos para descobrir tópicos de grandes coleções de documentos. Estes métodos são do tipo de aprendizagem não supervisionada. Assume que cada documento é composto por uma mistura de tópicos e que cada tópico representa uma distribuição de probabilidade das palavras (alguns podem ser sinónimos).

*Probabilistic Latent Semantic Analysis* (pLSA) (Hofmann, 1999) e Latent Dirichlet Allocation (LDA) (Blei et al., 2003) são as abordagens principais.

Ling et al. (2007) construíram um modelo de *aspect-sentiment mixture*. Este modelo foi baseado em pLSA e adquiriu uma aprendizagem com ajuda de alguns dados externos para a fase de treino.

A maior parte dos restantes modelos propostos são baseados em LDA. McDonald e Titov (2008) mostraram que os modelos de tópicos globais, tais como LDA podem não ser adequados para a deteção de aspetos. O motivo é que LDA depende das diferenças de distribuição dos tópicos e das coocorrências de palavras entre documentos. Os documentos de opinião sobre determinados produtos são bastante homogéneos, isto significa que em cada documento falam-se dos mesmos aspetos o que torna o modelo ineficaz. Este modelo apenas consegue ser eficaz na extração de entidades, como por exemplo *nomes de produtos* ou *nomes de marcas*. O modelo global descobre entidades enquanto o modelo local descobre aspetos usando algumas frases. Esta técnica não consegue separar aspetos e palavras de sentimento.

Ainda na abordagem não supervisionada foram implementados métodos de classificação baseados em grafos (Grineva et al. 2009), métodos de agrupamento (Liu et al., 2009) e métodos de modelação do idioma (Tomokiyo et al. 2003). *TextRank* (Mihalcea e Tarau, 2004) é uma das mais conhecidas abordagens de *graph-based approach*.

O método baseado em grafos (*graph-based approach*) consiste na construção de um grafo a partir de um documento de entrada. De seguida classificam-se os seus nós, de acordo, com a sua importância. Cada nó do grafo corresponde à palavra-chave candidata de um documento e uma aresta representa uma ligação entre dois candidatos relacionados. O peso da aresta é proporcional à relevância sintática ou semântica entre os candidatos ligados. A pontuação de um nó no grafo é definida de forma recursiva, em termos das arestas que tem e das pontuações dos nós vizinhos. Os candidatos mais bem classificados do grafo são, então, selecionados como palavras-chave para o documento de entrada.

Os métodos não supervisionados têm a vantagem de não exigir a construção de um *corpus* anotado.

## 2.4 Medidas de Avaliação

Quando os valores de sentimento são obtidos por avaliação humana representam o "*golden standard*", enquanto os outros valores são obtidos pelo sistema. Então, o objetivo principal da avaliação é comparar o valor do sentimento atribuído a um texto pelo sistema com o valor atribuído previamente por uma pessoa.

Existem diversas medidas usadas para avaliar classificadores, como por exemplo a *taxa de erro, precisão, recall, medida de F1*, entre outras. Todas estas medidas podem ser calculadas a partir da matriz de confusão, onde é possível observar a quantidade de acertos e erros para cada classe.

| | | **CLASSE PREDITA** | |
|---|---|---|---|
| **CLASSE ACTUAL** | | **P** | **N** |
| | **P** | Verdadeiro Positivo (VP) | Falso Negativo (FN) |
| | **N** | Falso Positivo (FP) | Verdadeiro Negativo (VN) |

**Figura 2.4** – Exemplo típico de uma matriz de confusão para duas classes.

**Verdadeiro Positivo** (acerto): Corresponde ao número de exemplos corretamente classificados como sendo da classe P.

**Falso Negativo** (erro): Indica o número de exemplos que são classificados como pertencentes à classe N, mas que verdadeiramente pertencem à classe P.

**Falso Positivo** (erro): Corresponde ao número de exemplos classificados como P, mas que pertencem à classe N.

**Verdadeiro Negativo** (acerto): Indica o número de exemplos corretamente analisados como pertencentes à classe N.

De seguida são apresentadas algumas medidas de desempenho:

**Taxa de erro**: Proporção do número de exemplos incorretamente classificados como pertencendo a uma classe que não sendo a sua no total de exemplos.

$$Taxa\ de\ erro = \frac{\mathrm{FP} + \mathrm{FN}}{\mathrm{VP} + \mathrm{VN} + \mathrm{FP} + \mathrm{FN}} \tag{2.6}$$

**Precisão**: Proporção de exemplos positivos classificados corretamente, entre todos os exemplos preditos como positivos.

$$Precisão = \frac{\text{VP}}{\text{VP} + \text{FP}} \tag{2.7}$$

**Revogação (*Recall*):** Proporção de exemplos positivos que foram classificados corretamente entre todos os exemplos positivos.

$$Recall = \frac{\text{VP}}{\text{VP} + \text{FN}} \tag{2.8}$$

**Medida F1**: É uma medida que combina a *precisão* e o *recall*.

$$Medida\ F1 = 2 * \frac{\text{Precisão} * Recall}{\text{Precisão} + Recall} \tag{2.9}$$

Para um problema que envolve duas classes podemos definir *precisão, recall* e *F1* em relação a uma classe, em que consideramos a classe positiva. Se por outro lado, considerarmos a outra como classe positiva, temos outros valores de *precisão, recall* e *F1*. A macro F1 combina os valores das duas classes.

**Macro F1**: É uma medida que calcula a eficácia da classificação, em que se parte do princípio que cada classe é igualmente importante. Por isso é que se atribuem pesos iguais às classes, independentemente da quantidade de documentos contidos. Através da fórmula em (2.10), é possível observar a combinação entre duas classes, $i$ e $j$ .

$$Macro\ F1 = 2 * \frac{\frac{(Precisão_i + Precisão_j)}{2} * \frac{(Recall_i + Recall_j)}{2}}{\frac{(Precisão_i + Precisão_j)}{2} + \frac{(Recall_i + Recall_j)}{2}} \tag{2.10}$$

## 2.5 Exemplos de aplicações práticas

Este estudo é focado para reclamações ou comentários de clientes. Estes podem ser analisados com recurso às técnicas de *Text Mining* e de análise de sentimentos, com o objetivo de extrair, por exemplo o tópico e o valor de sentimento. A figura seguinte mostra um exemplo de reclamação.


**Figura 2.5** – Reclamação via *email* de um cliente insatisfeito (Reproduzido de Ross Perez, 2013).

Todo este processo de *Text Mining* tem associado um dicionário de palavras comuns e a sua respetiva classificação (valor de sentimento). Com a construção de um léxico personalizado ou de palavras específicas relativas à área de negócio, cria-se uma ferramenta de valor acrescentado. O vocabulário específico é importante, pois permite analisar o *feedback* dos clientes. Existem várias situações de negócio que podem ser identificadas através do texto, como por exemplo *perda ou retenção de clientes, intenção de compra e venda*, *deteção de fraude*, entre outros. Na Figura 2.6 é mostrado um exemplo de um cliente que pretende efetuar uma compra.


**Figura 2.6** – Intenção de compra (Reproduzido de Alex Williams, 2012).

Uma equipa, da Universidade de Cornell, desenvolveu métodos automáticos sofisticados para detetar comentários falsos com base na extração de conhecimento a partir do texto, que pode ser visualizada na figura seguinte.

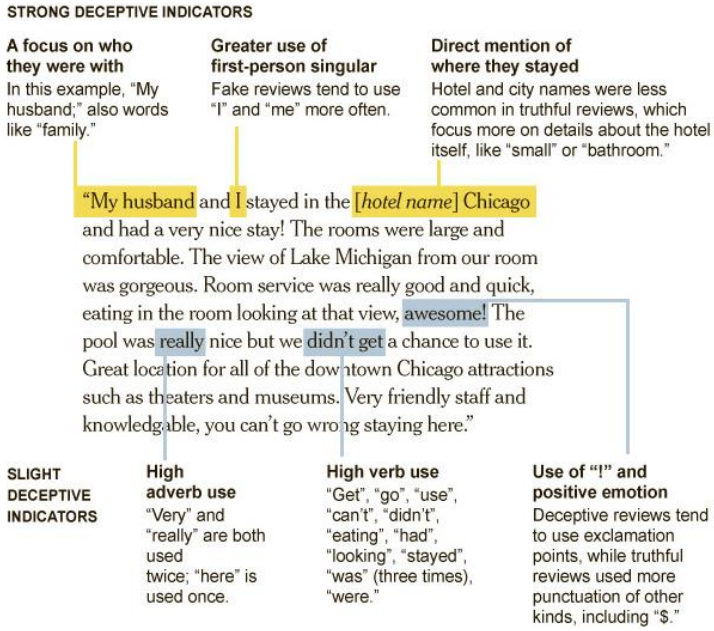



**Figura 2.7** – Características que o algoritmo procura (Reproduzido de David Streitfeld, 2011).

Os benefícios provenientes desta análise são preciosos para empresas que querem alcançar um patamar de excelência em determinadas vertentes de negócio, pois permitem aumentar a inteligência dos negócios e estar um passo à frente dos concorrentes. Atualmente, identificar informações "*escondidas*" nos dados é como encontrar "*diamantes*", como aliás o termo *Data Mining* sugere.

O próximo capítulo começa com uma motivação para a análise do caso de estudo. De seguida, descreve-se como é constituído o conjunto de dados e por último são mostrados todos os passos utilizados para a construção do sistema automático de classificação de sentimento de frases, onde são apresentados e analisados os resultados.

# CAPÍTULO 3

# ANÁLISE DE SENTIMENTOS DE COMENTÁRIOS DE UTILIZADORES: SISTEMA BÁSICO

Este capítulo mostra com algum detalhe todos os passos realizados ao longo desta tese. O ponto de partida para este trabalho são os comentários reproduzidos pelos assistentes de apoio ao cliente, os quais são disponibilizados por uma empresa de telecomunicações portuguesa. Todas as ferramentas aplicadas em torno desses comentários tem o objetivo de extrair informações relevantes. Estas ferramentas de apoio servem para identificar expressões de sentimento em relação a determinados assuntos. Inicialmente é explicado como é constituído o conjunto de dados e de seguida mostram-se todos os passos que conduziram à construção do sistema, bem como são discutidos e interpretados os seus resultados. Neste capítulo optou-se por descrever o sistema básico, sendo que no capítulo seguinte, será abordada a versão mais desenvolvida que apresenta melhor desempenho.

## 3.1 Motivação

Em determinadas indústrias existe um grande número de contactos e de interações com os clientes, quer através de *call centers* ou mesmo de escritórios de atendimento ao cliente. Por essa razão, estes dados não estruturados são a principal base para a análise e para assegurar uma boa tomada de decisão. Então, analisar o sentimento que o cliente transmite ao longo do tempo deve ser considerada uma prioridade, visto que se a empresa não conseguir atender ao seu pedido ou antecipar a resolução dos problemas, muito provavelmente o cliente acaba por abandonar os serviços dessa mesma empresa. Nos dias de hoje, perceber se o cliente está ou não satisfeito com os produtos e serviços tornou-se uma preocupação constante. Com esta análise as empresas podem desenvolver novas estratégias, que permitam modificar a forma como abordam os seus clientes e consequentemente auxiliar na sua retenção a longo prazo.

## 3.2 Recursos utilizados

A ferramenta utilizada nesta dissertação foi o *software* R (versão 3.0.3.).

R é uma linguagem de programação desenvolvida para criar cálculos estatísticos e gráficos. O facto de terem chamado R ao programa deveu-se em grande parte pelas iniciais do nome dos criadores começarem pela letra R, nomeadamente Ross Ihaka e Robert Gentleman. Este pode ser adquirido de forma gratuita no *site* CRAN (*The Comprehensive R Archive Network*), disponível em http://www.r-project.org/. É um recurso bastante poderoso, onde se podem utilizar pacotes que contêm funções ou áreas de estudo específicas, com o intuito de apoio à programação.

## 3.3 Descrição de dados

A coleção de documentos contém comentários datados desde 1 de Junho de 2014 até 3 de Setembro de 2014, perfazendo um total de 1724 casos.

O *dataset* é composto por três variáveis e os dados apresentam-se sob a forma de uma tabela. A primeira variável é a *data*, na qual está incluída o dia e a hora a que são contactados; a segunda variável é a *identificação do cliente* que se encontra codificada através de letras e números, e por último têm-se os *comentários dos assistentes* em forma duma breve descrição da mensagem que o cliente lhe transmitiu, em tempo real.

Através da Tabela 3.1, pode-se observar uma pequena amostra para compreender um pouco melhor como é a estrutura dos dados.

| Date | TOPIC_ID | Notes |
|---|---|---|
| **01/06/2014 08:12** | C26E5… | informei que os minu sao creditados durante o dia 1 |
| **01/06/2014 09:31** | A33E4… | Clt solicitou barramento total de sva's |
| **01/06/2014 08:28** | 5B9D7… | informei que os minutos seriam creditados durante o dia de hoje |

**Tabela 3.1** – Pequena amostra do conjunto de dados.

Estes documentos de texto resultam do contacto do cliente com o assistente ou vice-versa, com o objetivo de discutir algum assunto pertinente.

É de esperar, que a maior parte dos comentários estejam escritos em português, uma vez que a empresa é portuguesa. À partida, pode-se detetar que existe uma desvantagem em relação à língua utilizada, pois os recursos disponíveis são escassos e limitados quando comparados com os que existem para a língua inglesa. Contudo, dado que a língua portuguesa tem características bastante peculiares, a aplicação de algumas técnicas pode ser encarada como um desafio.

## 3.4 Como o sistema funciona: um exemplo

Neste sentido, é proposta uma metodologia para o estudo do sentimento dos clientes.

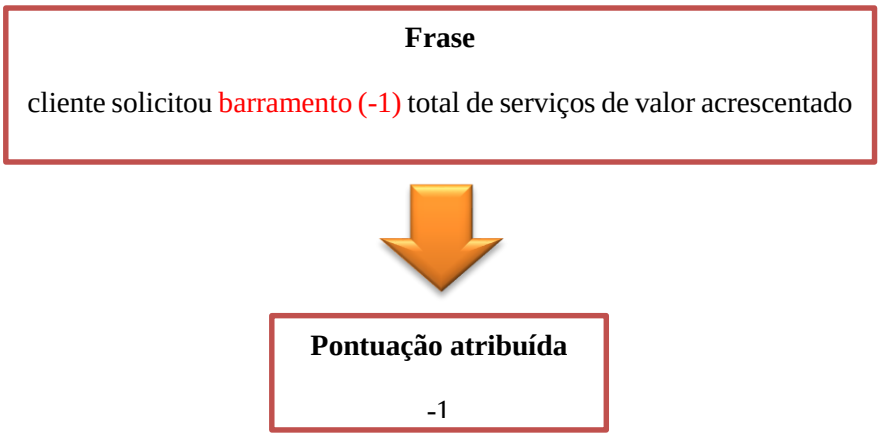



**Figura 3.1** – Proposta da metodologia com sistema básico para identificar o sentimento da frase.

Na Figura 3.1, pode-se constatar que esta frase diz respeito a um cliente que telefonou para a empresa e foi criado um registo com a informação relevante. Esta frase tem associado um sentimento negativo devido à palavra *barramento*. Tendo em consideração um léxico é atribuída a esta frase uma pontuação de -1.

Esta situação representa um cliente que quer abandonar determinados serviços, o que pode indicar algum perigo de perda de um cliente, existindo assim uma necessidade de estar mais atento aos próximos passos desse cliente.

Em conclusão, o que se pretende é identificar, de forma automática, clientes que poderão estar em risco de abandonar a empresa e que serão supostamente classificados como -3, bem como clientes que desejam adquirir mais serviços, sendo classificados como 3.

## 3.5 O sistema básico de análise de sentimentos

O sistema básico é composto por um conjunto de etapas habituais, que irão ser descritas daqui em diante, utilizadas para desenvolver todo o processo de *Text Mining* e de Análise de Sentimentos. Este sistema é o início e a base de todo o projeto, uma vez que inclui os passos fundamentais para a elaboração deste trabalho.

Para esta dissertação utilizou-se a abordagem baseada em dicionários. Esta abordagem consiste no uso de um léxico pré-definido que contém palavras positivas e negativas. De seguida, as frequências das palavras são calculadas para posteriormente ser possível pontuar todas as opiniões do conjunto de dados.

A grande dificuldade neste trabalho incidiu, sobretudo, no facto do conjunto de dados estar em português. Isto porque a maior parte dos estudos efetuados são para a língua inglesa, como já foi referido anteriormente. Logo, a procura por um léxico que estivesse em português foi bastante árdua. Então, para concretizar este estudo fez-se recurso a dois léxicos já existentes: o léxico Afinn e o léxico Sentilex.

### 3.5.1 Léxico Afinn

O léxico Afinn é um léxico com 2477 palavras e frases em inglês, com uma escala de números inteiros que vai de -5 a 5, para a qual a classificação de -5 está associada a palavras mais negativas, enquanto a classificação de 5 está associada a palavras com caráter positivo. As palavras foram marcadas manualmente por Finn Nielsen entre 2009 a 2011.

| | | |
|---|---|---|
| advanced | → | 1 |
| advantage | → | 2 |
| advantages | → | 2 |
| adventure | → | 2 |
| adventures | → | 2 |
| adventurous | → | 2 |

**Figura 3.2** – Pequena amostra do léxico Afinn.

Para poder identificar esta lista de palavras nos textos, este léxico Afinn foi traduzido para português, de forma manual, permanecendo 1997 palavras.

| | | |
|---|---|---|
| avançado | → | 1 |
| vantagem | → | 2 |
| vantagens | → | 2 |
| aventura | → | 2 |
| aventuras | → | 2 |
| aventureiro | → | 2 |

**Figura 3.3** – Pequena amostra do léxico Afinn traduzida manualmente.

Este léxico ficou com uma lista de vocabulário mais reduzida, devido ao facto de algumas palavras em inglês, bem como frases não terem tradução literal para a língua portuguesa. Foi um processo bastante demorado uma vez que teve que ser analisada palavra a palavra.

### 3.5.2 Léxico Sentilex

Relativamente ao léxico Sentilex, é um léxico em português com 82.347 formas flexionadas, ou seja, contém expressões de apenas uma única palavra (expressões simples) ou com mais de uma palavra (expressões compostas) e também expressões idiomáticas típicas portuguesas. Este léxico assume uma escala de números inteiros que varia entre -1 a 1. Este recurso foi desenvolvido precisamente para aplicações de extração e de classificação de sentimentos em português (Silva et al., 2012).

| | | |
|---|---|---|
| ativo | → | 1 |
| inaceitável | → | -1 |
| inacessível | → | -1 |
| inacreditável | → | 0 |
| inativa | → | -1 |
| inatividade | → | -1 |

**Figura 3.4** – Pequena amostra do léxico Sentilex.

Com base nestes dois léxicos, Afinn e Sentilex, o que vai ser feito é o reconhecimento destas palavras no conjunto de dados para que possam ser calculadas as pontuações das frases. As palavras que não são identificadas por estes dois léxicos são omitidas nesta

análise, ou seja, são palavras que os léxicos não conseguem cobrir pois não existem lá. Este modelo básico será realizado para ambos léxicos, de forma individual, para depois ser possível avaliar e obter o seu desempenho.

### 3.5.3 Método de pré-processamento

O conjunto de etapas deste sistema envolve a identificação destas palavras nos léxicos e no texto, o pré-processamento habitual, o cálculo do valor de sentimento para cada frase, a conversão desse valor para uma escala de números inteiros que vai desde -3 a 3 e a análise do custo de cada frase. Estas etapas vão ser abordadas ao longo deste capítulo.

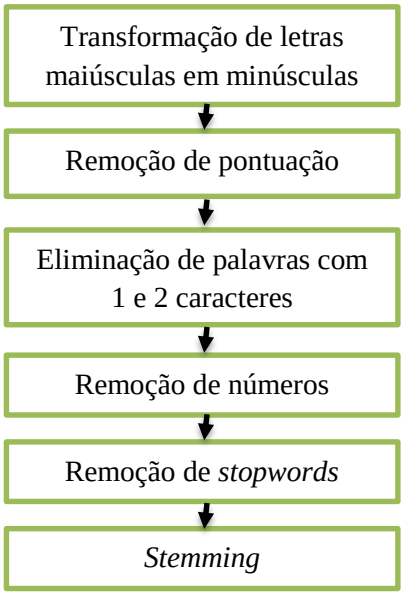



**Figura 3.5** – Etapas do pré-processamento habitual.

Através desta figura, é possível observar todas as etapas que incluem o pré-processamento habitual.

A primeira etapa remete para a transformação de letras maiúsculas em minúsculas para que seja possível identificar palavras que são iguais, palavras essas que não eram identificadas devido a conterem pelo menos uma letra maiúscula como, por exemplo, seria o caso de uma palavra ao iniciar uma frase.

Relativamente às seguintes etapas que compreendem a remoção de números, pontuação, eliminação de palavras com 1 e 2 caracteres, espaços em branco, bem como a eliminação

de *stopwords*, são etapas que permitem eliminar termos não relevantes e que transmitem pouca informação para este estudo.

Por último tem-se o *stemming* que apaga o sufixo da palava ficando apenas o seu radical. Uma vez que os algoritmos de *stemming* dependem estritamente das regras de formação de palavras e como já foi dito no Capítulo 2, que existem alguns problemas na aplicação em outras línguas, optou-se por utilizar este processo de forma ponderada.

Pois imaginemos que se aplicasse o *stemming* ao texto todo, muito provavelmente haveriam palavras que poderiam ter o mesmo radical mas com pontuações diferentes, o que seria um grave problema.

| **Léxico** | **Palavras** | **Radical** | **Valor** |
|---|---|---|---|
| Afinn | afetado | → afet | → -2 |
| Afinn | afeto | → afet | → 3 |
| Sentilex | falar | → fal | → 0 |
| Sentilex | falido | → fal | → -1 |

**Figura 3.6** – Palavras que possuem o mesmo radical

Através desta figura pode-se constatar o que foi dito anteriormente. Precisamente por esta razão, nesta dissertação, o *stemming* não é aplicado ao texto todo, mas sim é apenas aplicado quando uma palavra do texto não esteja a ser identificada no respetivo léxico. Dessa forma, faz-se o stemming dessa palavra no texto e descobre-se o seu valor de sentimento no léxico.

| **Texto original** | **Texto após pré-processamento** |
|---|---|
| "cliente **no** dia foi descontado **o** acesso **a** internet **no** entanto **a** cliente menciona **que** não usou nem activou pede **a** devolução **do** saldo retirado **e o** bloqueio **da** internet" | cliente dia foi descontado acesso internet entanto cliente menciona não usou nem activou pede devolução saldo retirado bloqueio internet" |
| "cliente pretende desactivar todos **os** valores extra **ao** saldo **por esta a** ser cobrado **4 por** semana**.**" | "cliente pretende desactivar todos valores extra saldo ser cobrado semana" |

**Tabela 3.2** – Transformação do texto com pré-processamento habitual.

Notemos que nesta tabela estão compreendidas as etapas do pré-processamento habitual, sendo bastante visível nestes exemplos a remoção de números, de *stopwords* e de pontuação.

### 3.5.4 Uniformização de valores

Uma vez que a escala dos dois léxicos é diferente, o passo seguinte consiste em converter as escalas dos dois léxicos para uma escala de números inteiros que varia entre -3 e 3.

Para tal, foi feito o somatório de cada frase para os respetivos léxicos e observou-se o valor mínimo e máximo resultante dessa operação. Para poder tomar a escala desejada, divide-se cada frase por um número, cujo seu valor é obtido da seguinte forma. Em primeiro lugar, o máximo ou mínimo é dividido por 3. Assim, o somatório correspondente a cada frase é dividido pelo valor resultante da operação anterior, ficando desta maneira a escala compreendida entre -3 e 3.

Após várias tentativas para descobrir qual o valor que melhor se adequava a cada léxico, foi estabelecido o valor de 8/3 para o léxico Afinn e o valor de 1 para o léxico Sentilex. Uma vez concretizado o passo anterior é feito um arredondamento para que os valores assumam uma escala discreta.

## 3.6 Resultados e análise

O primeiro passo consiste em avaliar e interpretar os resultados dos dois léxicos: Afinn e Sentilex. A *performance* destes dois léxicos é feita através das medidas de desempenho descritas no Capítulo 2.

Em primeiro lugar, são apresentados alguns exemplos de frases classificadas pelos léxicos, de seguida é colocada uma tabela com os respetivos valores de sentimento obtidos pelo sistema básico e por último são apresentados os histogramas com a distribuição dos resultados.

Através da Tabela 3.3, é possível observar que a vermelho estão representadas as palavras negativas e a verde as palavras positivas pertencentes ao léxico respetivo, bem como está indicado o somatório final das pontuações de cada palavra.

| **Léxico** | **Frase** | **Valor (Somatório)** |
|---|---|---|
| AFINN | 1453: "**resolvido (2)** internet funcionar" | → 2 |
| AFINN | 504: "cliente **reclama (-2)** facto compra telemovel pontos loja **não (-1)** foi *atribuida* oferta semanas internet *possivel* efetuar atribuição" | → -3 |
| SENTILEX | 166: "cliente loja reclama *fato* ter aderido serviço ainda *nao* estar **ativo (1)** contrato efetuado loja" | → 1 |
| SENTILEX | 244: "cliente loja reclama *apos* receber proposta *atraves* mudança tarifário informaram manter altura tinha **direito (1)** neste momento não tem acesso nada peço verifiquem situação *pff*" | → 1 |

**Tabela 3.3** – Exemplos de frases classificadas pelos léxicos Afinn e Sentilex

Podemos constatar que na maior parte dos casos apenas uma palavra é classificada na frase e que as duas últimas frases não estão bem classificadas. Com o léxico Sentilex é possível notar, que só foram encontradas palavras positivas como "*ativo*" e "*direito*", como tal a pontuação refletida também é positiva. Contudo, se analisarmos as frases no geral, vê-se que tem bastantes palavras negativas como é o caso de "*reclama*", "*não*", entre outras. Também se observa que algumas frases contém erros ortográficos e algumas abreviaturas como por exemplo "*nao*", "*pff*", entre outras. O próximo capítulo mostra como estes exemplos são tratados.

### 3.6.1 Avaliação de classificações

Para avaliar os resultados foi efetuada uma análise com base no custo de cada frase que indicará qual será a melhor tomada de decisão, ou seja, a decisão de manter ou eliminar as etapas que foram referidas anteriormente neste capítulo.

| | *-3* | *-2* | *-1* | *0* | *1* | *2* | *3* |
|---|---|---|---|---|---|---|---|
| *-3* | **0,00** | 0,17 | 0,33 | 0,50 | 0,67 | 0,83 | **1,00** |
| *-2* | 0,17 | **0,00** | 0,17 | 0,33 | 0,50 | 0,67 | 0,83 |
| *-1* | 0,33 | 0,17 | **0,00** | 0,17 | 0,33 | 0,50 | 0,67 |
| *0* | 0,50 | 0,33 | 0,17 | **0,00** | 0,17 | 0,33 | 0,50 |
| *1* | 0,67 | 0,50 | 0,33 | 0,17 | **0,00** | 0,17 | 0,33 |
| *2* | 0,83 | 0,67 | 0,50 | 0,33 | 0,17 | **0,00** | 0,17 |
| *3* | **1,00** | 0,83 | 0,67 | 0,50 | 0,33 | 0,17 | **0,00** |

**Tabela 3.4** – Representação da tabela de custos.

A tabela de custos foi divida em 7 patamares para obter um custo máximo de 1 e para, desta forma, podermos comparar os resultados obtidos na base de custos com as taxas de erro. Na diagonal estão os casos que têm custo zero, ou seja, são os casos que foram corretamente classificados em cada classe. À medida que existe um afastamento da diagonal, o custo vai aumentando, sendo que nos extremos é igual a 1. Isto facilmente é explicado, pois o custo de classificar um exemplo como -3 quando na realidade tem um valor de sentimento igual a 3 é bem superior ao custo de classificar uma frase como -2 e na realidade esta ser um -3. Para procedermos ao cálculo do custo total médio por frase, multiplicou-se a matriz de confusão pela matriz de custos unitária. De seguida, dividiu-se a matriz resultante da operação anterior pelo número total de casos estudados e por fim somamos todos os valores.

### 3.6.2 Resultados com léxico Afinn

Através da Tabela 3.5, é possível observar os resultados para o sistema básico com o léxico Afinn, onde são utilizadas 4 variantes. A primeira variante diz respeito à introdução do léxico Afinn, em que o sistema irá identificar somente as palavras que estiverem contidas neste léxico. A segunda variante remete para a inclusão da primeira variante juntamente com o pré-processamento habitual, em que o texto sofre algumas modificações, descritas em 3.4.3, para que seja possível a identificação de mais palavras. O terceiro passo consiste na junção das duas variantes anteriores em conjunto com a eliminação de *stopwords*. Por último, tem-se todas as variantes atrás mencionadas adicionando o *stemming*. Nesta tabela também são usadas duas medidas de avaliação: a taxa de acerto e o custo médio.

| | | | | | **Léxico Afinn** | |
|---|---|---|---|---|---|---|
| **Variantes** | **1** | **2** | **3** | **4** | **Taxa de acerto %** | **Custo Médio** |
| **1** | x | | | | 33,83 | 0,1685 |
| **2** | x | x | | | 35,46 | 0,1635 |
| **3** | x | x | x | | 35,46 | 0,1635 |
| **4** | x | x | x | x | **38,45** | **0,1610** |

**Tabela 3.5** – Resultados com metodologia do sistema básico com o léxico Afinn.

Identifica-se que a taxa de acerto vai crescendo gradualmente nas 4 variantes da tabela, contudo a variante que obteve melhor desempenho foi o *stemming*. Quanto ao custo

médio por frase, observa-se também uma diminuição ao longo de todas as variantes, sendo a quarta variante que apresenta um menor custo médio por frase.

No caso da segunda variante, os resultados melhoram tanto a nível da taxa de acerto como a nível de custo médio por frase. Muito provavelmente isto significa que o sistema está a encontrar mais palavras e consequentemente está a alterar as pontuações das frases de forma positiva.

A eliminação de *stopwords* não altera em nada os resultados, pois são palavras que têm pouca ou nenhuma informação e que nem sequer estão pontuadas pois não faria sentido. A lista destas palavras poderá ser encontrada no Anexo A.

Quanto ao *stemming*, como já foi dito assiste-se a um aumento da taxa de acerto, o que significa que os casos bem classificados em cada classe da matriz de confusão estão a aumentar. Como apenas se procura o radical da palavra no léxico, isto significa que pelo menos o sistema está a conseguir identificar mais palavras que não constavam anteriormente no léxico. O custo médio por frase, nesta variante, mostra um decréscimo, pois os casos mal classificados em cada classe poderão estar menos pronunciados, ou seja, existem erros menores. Tem-se a título de exemplo frases que foram classificadas como -3 quando na verdade deveriam ser classificadas como -2.

### 3.6.3 Resultados com léxico Sentilex

Para a análise dos resultados do léxico Sentilex foram também admitidas as 4 variantes e as duas medidas de desempenho, que foram devidamente explicadas na secção anterior.

| | | | | | **Léxico Sentilex** | |
|---|---|---|---|---|---|---|
| **Variantes** | **1** | **2** | **3** | **4** | **Taxa de acerto %** | **Custo Médio** |
| **1** | x | | | | 30,70 | 0,1873 |
| **2** | x | x | | | **31,46** | 0,1923 |
| **3** | x | | x | | 30,70 | 0,1873 |
| **4** | x | | x | x | 30,70 | **0,1873** |

**Tabela 3.6** – Resultados com metodologia do sistema básico com o léxico Sentilex.

Através desta tabela, observa-se que a taxa de acerto apenas aumenta com o pré-processamento habitual, permanecendo constante nas outras variantes. O custo médio por

frase aumenta na segunda variante para 0.1923, tomando como melhor valor 0.1873 nas restantes variantes.

No caso do léxico Sentilex observa-se um pequeno agravamento nos resultados com o pré-processamento habitual. Isto deve-se ao facto do sistema estar a encontrar mais palavras, o que neste caso não é necessariamente bom. Uma vez que este léxico toma uma escala de valores inteiros entre -1 a 1, as palavras encontradas anulam-se umas às outras e consequentemente as pontuações das frases são alteradas.

Quanto à terceira variante, a eliminação de *stopwords*, não representa um impacto significativo nos resultados pelo mesmo motivo descrito anteriormente em 3.5.2.

Por último tem-se o *stemming* que também não influencia os resultados. Isto significa que muito provavelmente não foram encontradas mais palavras. Este problema é explicado pelo facto deste léxico apresentar uma taxa de cobertura de palavas de apenas 5%, o que é extremamente baixa.

As matrizes de confusão e os gráficos de custo dos léxicos, Afinn e Sentilex, são um pouco extensas e por esse mesmo motivo poderão ser vistas no Anexo B e C.

### 3.6.4 Comparação de resultados dos dois léxicos

Estes histogramas mostram a distribuição dos resultados para cada léxico e foram construídos através dos resultados anteriores, neste capítulo. A última variante considerada na construção do histograma inclui o *stemming* para ambos léxicos.

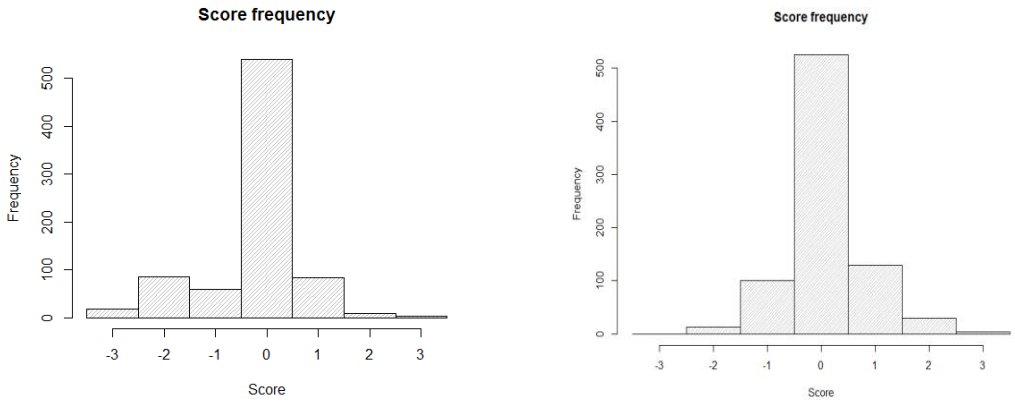



**Figura 3.7** – Histograma com distribuição de resultados dos léxicos Afinn (à esquerda) e Sentilex (à direita).

Nestes histogramas estão representados os 801 textos dos assistentes. Também são apresentadas as frequências de valores previstos para cada classe como, por exemplo, para o léxico Afinn são obtidas 86 previsões com valor -2 e para o léxico Sentilex são obtidas 110 previsões com valor 1.

### 3.6.5 Reflexões sobre os resultados obtidos

Se analisarmos atentamente as taxas de acerto de ambos os léxicos vemos que o léxico Afinn ronda os 38% com um custo médio por frase de 0,1610 e o léxico Sentilex ronda os 30% com um custo médio de 0,1873. Então, observando este panorama gera-se a pergunta de "como posso melhorar estes resultados?" Claramente, o sistema básico apresenta algumas falhas que precisam de ser melhoradas.

Uma das melhorias que pode trazer grandes benefícios poderá ser a criação de um léxico enriquecido com vocabulário específico desta unidade de negócio, uma vez que os dois léxicos, Afinn e Sentilex, são demasiado gerais. Na área de telecomunicações, as reclamações são muito importantes, visto que a satisfação do cliente está acima de tudo. Notemos que no léxico Afinn a palavra "*reclama*" está pontuada como -2 e no léxico Sentilex não existe. Perante esta situação, parece que modificar a sua relevância e introduzir no léxico esta palavra, respetivamente, não seria uma má ideia. E como esta palavra existem muitas outras que merecem esta atenção.

Existem frases pré-definidas ditas pelos operadores de telecomunicações que podem ser consideradas como *stopwords*, pois são bastante comuns e não acrescentam qualquer informação relevante, como por exemplo "*Boa tarde, bem-vindo ao apoio ao cliente se pretende ajuda sobre um tema (...)*". Neste caso as palavras "*boa*", "*tarde*", "*bem*", "*apoio*", "*ajuda*" iriam ser classificadas o que não faria sentido, pois são palavras que não transmitem sentimento e que iriam influenciar as pontuações finais das frases.

Outro problema identificado é o facto de frases negativas não refletirem essa negatividade. A palavra "*não*" apenas consta no léxico Afinn e está classificada como -1. Evidentemente, com uma pontuação tão pouco relevante, o que vier a seguir na frase facilmente é contrariado.

Palavras como "*muito*", "*demasiado*" amplificam o sentimento, por outro lado palavras como "*pouco*", "*menos*" desamplificam o sentimento. Na língua portuguesa existe a necessidade de causar, propositadamente, este impacto nas frases, pois é uma língua que apresenta bastantes peculiaridades. Tal como o caso anterior, esta situação também não foi considerada neste sistema básico.

Outro problema bastante visível é também os operadores escreverem com alguns erros ortográficos como foi visto nos exemplos anteriores, ou seja, palavras como "*nao*", "*atribuicao*", "*possivel*", entre outras não são identificadas e como tal não são pontuadas.

Por estes motivos, entre tantos outros, fazem com que o sistema básico não tenha um bom desempenho, nem bons resultados. Por isso, surge a necessidade de criar um modelo avançado que permitirá alcançar resultados mais robustos e significativos.

# CAPÍTULO 4

# ANÁLISE DE SENTIMENTOS DE COMENTÁRIOS DE UTILIZADORES: SISTEMA AVANÇADO

Este capítulo começa por mostrar todas as alterações que foram introduzidas ao sistema básico, o qual será denominado de *sistema avançado*. Este sistema avançado pretende ser uma extensão do sistema básico. O objetivo da implementação deste novo método é que traga algumas melhorias ao sistema anterior.

Neste capítulo pretende-se apresentar os resultados do sistema avançado, bem como incluir um léxico otimizado para cada base de dados. Por último, para validar o sistema avançado, este será comparado com métodos de aprendizagem supervisionada. Seguidamente, estes resultados são analisados com o objetivo de poder melhorá-los.

## 4.1 Metodologia com sistema avançado

O sistema avançado foi criado devido à necessidade de obter melhores resultados. Este novo sistema tem como base o sistema básico, ou seja, todas as etapas mencionadas no Capítulo 3 foram utilizadas para o seu desenvolvimento. De seguida, apresentamos todas as etapas usadas na construção deste sistema.

### 4.1.1 Léxicos

É preciso ter em conta que utilizar um léxico geral, normalmente, não produz bons resultados. Quando o léxico geral é combinado com um léxico da área de interesse, o sistema fornece melhores resultados.

Por esse motivo foi construído de raiz um léxico enriquecido, *DomainWords*, utilizando como base o léxico Afinn e a sua escala de números inteiros que varia entre -5 a 5.

| | | | | | |
|---|---|---|---|---|---|
| desativou | → | -5 | desativou | → | -1 |
| desesperado | → | -3 | desesperado | → | -1 |
| desistência | → | -5 | desistência | → | -1 |
| interessado | → | 5 | interessado | → | 1 |
| satisfeita | → | 4 | satisfeita | → | 1 |

**Figura 4.1** – Amostra do léxico enriquecido, *DomainWords*, representada à esquerda com a escala de-5 a 5 e à direita com a escala -1 a 1.

O léxico *DomainWords* conta com 6.915 palavras introduzidas manualmente, onde está contido o vocabulário inerente à nossa área das telecomunicações. Este processo demorou uns largos meses. Algumas das pontuações foram ajustadas tomando em consideração o facto de que alguns termos nesta área serem muito pejorativos.

### 4.1.2 Pré-processamento

No sistema avançado reutilizamos todas as etapas do pré-processamento habitual, descritas anteriormente no Capítulo 3. No entanto, introduzimos uma novidade o pré-processamento avançado que compreende um conjunto de três etapas que são explicadas seguidamente.

A primeira etapa consistiu na alteração de algumas abreviaturas para a sua forma normalizada, caso ilustrado na Figura 4.2 e na correção de palavras com erros ortográficos, representado pela Figura 4.3.

| | | |
|---|---|---|
| cl | → | cliente |
| cli | → | cliente |
| clie | → | cliente |
| clt | → | cliente |
| clte | → | cliente |

**Figura 4.2** – Lista de diversas variantes da palavra "*cliente*".

Normalmente, nos textos escritos pelos operadores, encontram-se palavras que podem ter várias variantes ao longo dos documentos de texto.

| | | | | | |
|---|---|---|---|---|---|
| ic | → | incumprimento | ameça | → | ameaça |
| tccs | → | tentativa de contato com sucesso | nao | → | não |
| tcss | → | tentativa de contato sem sucesso | adesao | → | adesão |
| chm | → | chamada | decrescimo | → | decréscimo |
| cl | → | cliente | necesasrias | → | necessárias |

**Figura 4.3** – Exemplos de algumas regras de reescrita.

Para resolver este problema foi construída uma lista com 635 regra de reescrita, para que houvesse um reconhecimento destes termos nos léxicos. Estas palavras estão com as duas formas, ou seja, com o antigo e novo acordo ortográfico, pois os operadores escrevem das duas maneiras.

A segunda etapa, do pré-processamento avançado, compreende a remoção dos textos em que apenas está incluída uma palavra, pois uma palavra representa pouca ou mesmo nenhuma informação.

| **Léxico** | **Comentários** | **Valor Final** |
|---|---|---|
| AFINN | …: "**desativo (-2)**" | → 0 |
| SENTILEX | …: "**ativei (1)**" | → 0 |
| AFINN / SENTILEX | …: "**asssssdss**" | → 0 |

**Tabela 4.1** – Exemplo de remoção de textos em que está incluída uma única palavra.

Através do exemplo da Tabela 4.1, observa-se que se numa nota apenas estiver a palavra "*desativo*", assumimos que esta palavra não representa informação relevante, pois não existe mais informação acerca do assunto. É muito provável que a palavra que esteja sozinha na nota tenha uma pontuação, mas como não é informativa é removida.

A terceira etapa incide na construção de uma lista de frases e palavras bastante comuns, denominada de *stopwords personalizadas (my_stopwords)*. Existem frases pré-definidas ou gravadas pelos operadores que não revelam qualquer sentimento.

| **Léxico** | **Frase** | **Valor (Somatório)** |
|---|---|---|
| AFINN | …: "***Bem (2)*** -vindo ao ***apoio (2)*** ao cliente. ***Boa (3)*** ***tarde (-2)***, em que posso ser ***útil (2)***?" | → 7 |

**Tabela 4.2** – Exemplo de frase gravada pelos operadores.

Através do exemplo ilustrativo da Tabela 4.2, pode-se constatar que palavras como "*bem*", "*apoio*", "*boa*", "*tarde*", "*útil*", iriam ter associado um valor de sentimento, quando na realidade não deveriam.

### 4.1.3 Tratamento da Negação

Depois de acabadas todas as fases do pré-processamento, considerou-se que o impacto da negação nas frases deveria ser destacado e devidamente estudado.

A palavra "*não*", não é considerada no léxico Sentilex e no léxico Afinn apenas tem um valor de sentimento de -1, o que faz com que uma frase que à partida é muito negativa se torne positiva.

| **Léxico** | **Frase** | **Valor** |
|---|---|---|
| AFINN | 1159: "cliente **não (-1) consegue (2)** enviar mms outras redes depois reconfigurar serviço nem trocar equipamento peço *pfv verifquem*!" | → 1 |
| SENTILEX | 1265: "*ccs* cliente *não (-1) aceitou (0)* ofertas *retenção (0)* procedi *desativação (0)*" | → 0 |

**Tabela 4.3** – Amostra de casos em que a negação não é refletida com sistema básico.

Nos exemplos da Tabela 4.3, classificados com o sistema básico, é possível observar que para o léxico Afinn apenas duas palavras têm pontuação, ou seja, "*não*" que tem valor de -1 e "*consegue*" com um valor de 2. Fazendo o somatório das palavras vê-se que a pontuação final da frase é 1.

Com o léxico Sentilex nenhuma palavra foi identificada no exemplo, incluindo as palavras "*não*", "*aceitou*", "*retenção*", "*desactivação*", uma vez que estas palavras não existem neste léxico.

Como se pode observar, ambas as frases são extremamente negativas e os léxicos não estão a corresponder às expectativas.

Perante tal situação, elaboramos uma lista com 33 palavras que contrariam o sentido das frases para tentar solucionar o problema.

| | | |
|---|---|---|
| jamais | → | -1 |
| mas | → | -1 |
| não | → | -1 |
| nem | → | -1 |
| nunca | → | -1 |

**Figura 4.4** – Amostra da lista de palavras contidas na negação.

Nesta área de negócio, a grande parte das pessoas frequentemente liga para reclamar sobre alguma coisa com que estejam descontentes. Assim, o que se pretende é rearranjar a formulação da negação, de modo a que esta esteja refletida nos exemplos tratados. Em consequência foram consideradas duas hipóteses.

A primeira hipótese denominada de "*negação simples*", pois o seu intuito é que sempre que o sistema encontre no texto alguma das palavras incluídas na lista da negação, irá multiplicar a palavra seguinte por -1.

A segunda hipótese é igual à primeira, exceto, no caso em que a palavra seguinte à negação tem classificação negativa. Neste caso, a multiplicação por -1 fica sem efeito. Apenas o valor da palavra negativa é considerado. Para os restantes casos em que a palavra seguinte à negação é positiva ocorre a multiplicação por -1, tal como na primeira hipótese.

| **Léxico** | **Frase** | **Hipótese 1** | **Hipótese 2** |
|---|---|---|---|
| AFINN | 130: "cliente veio loja pedir **activação (1)** roaming contratual serviço ***não(-1)*** está **activo (2)** cliente muito **descontente (-3)**" | → -4 | → -4 |
| SENTILEX | 609: "cliente contactado informado vez mais reiterei preço (…) informei cliente dirigir loja informa ***não (-1)*** quer ser mais **incomodado (-1)** *não (-1)* quer receber mais contactos (…)" | → 1 | → -1 |

**Tabela 4.4** – Hipóteses consideradas na negação.

Na primeira frase está presente a hipótese 1 da negação, que contém a palavra "*não*" que apresenta um valor de sentimento igual a -1 e que irá inverter o valor da palavra positiva "*activo*". Fazendo o somatório da frase observa-se que o resultado final é -4. O mesmo se passa com a hipótese 2 da negação para essa mesma frase.

Na segunda frase, a hipótese 1 da negação também é aplicada, pois a palavra negativa "*incomodado*" é multiplicada por -1, trocando o sentido do que supostamente se pretendia. Contudo, na hipótese 2 da negação, o "*não (…) incomodado*" será ligeiramente diferente, pois "*incomodado*" é uma palavra negativa. Logo, apesar de estar o "*não*" atrás, a multiplicação não irá ter influência, permanecendo apenas o valor da palavra "*incomodado*".

### 4.1.4 Palavras Amlificadoras e Atenuadoras

De seguida, ponderou-se a utilização de palavras amplificadoras e atenuadoras do sentimento. Estas palavras fazem todo o sentido dado que se pretende estudar o sentimento que o cliente exprimiu.

Outro fator que enaltece esta questão é o facto de este estudo estar a ser feito para a língua portuguesa, aumentando o interesse desta questão, pois a língua portuguesa é bastante minuciosa. Por exemplo, se alguém disser "*gosto de chocolate*" não está a transmitir o mesmo sentimento como quando diz "*gosto muito de chocolate*", ou mesmo quando diz "*gosto muito pouco de chocolate*". São precisamente estas questões que se pretendem estudar.

Para resolver a situação, organizou-se uma lista com 166 palavras amplificadoras e com 35 palavras atenuadoras.

| | |
|---|---|
| bastante | dificilmente |
| demasiado | lentamente |
| imediatamente | menor |
| imediato | pouco |
| muito | quase |

**Figura 4.5** – Amostra de palavras amplificadoras (à esquerda) e atenuadoras (à direita).

### 4.1.5 Uniformização de valores

Para a conversão da escala de pontuações dos léxicos, procedeu-se de forma análoga à descrita no Capítulo 3, introduzindo pequenas modificações.

O valor estabelecido para o léxico Afinn foi de 4 e para o léxico Sentilex foi novamente 1, tendo como objetivo uma escala entre -2.5 e 2.5.

A razão de considerar esta escala é a introdução das palavras amplificadoras e atenuadoras, que a amplificam, ficando esta compreendida entre -3 e 3. O processo de cálculo é explicado na Figura 4.6.

Repara-se que a escala considerada neste processo, entre -2.5 e 2.5, não compromete as frases sem a presença de palavras amplificadoras e desamplificadoras, uma vez que com o arredondamento final, as frases com a classificação extrema (-2.5 e 2.5) tomam o valor de -3 e 3.

Se o **valor de sentimento** da frase for **diferente de zero**:

- Caso não existam amplificadores ou atenuadores o resultado mantêm-se.
- Caso existam amplificadores:
  - Se o valor da frase for positiva, adiciona-se 1.
  - Se o valor da frase for negativa, subtrai-se 1.
- Caso existam atenuadores:
  - Se o valor da frase for positiva, subtrai-se 1.
  - Se o valor da frase for negativa, adiciona-se 1.

Se o **valor de sentimento** da frase for **igual a zero**:

- Caso existam amplificadores, soma-se a classificação das palavras que têm amplificadores associados.
  - Se a soma das palavras for positiva, adiciona-se 1.
  - Se a soma das palavras for negativa, subtrai-se 1.

**Figura 4.6** – Cálculo da pontuação das frases com sistema avançado.

Com esta parte concluímos a descrição do sistema avançado. Na próxima secção descrevemos como o sistema foi avaliado.

## 4.2 Resultados e análise

Foram anotados manualmente 801 casos para posteriormente ser possível avaliar a fiabilidade dos resultados dos léxicos. Desta forma, fizeram-se umas experiências para ver até que ponto as diferentes facetas do sistema melhoravam ou não os resultados.

| **Léxico** | **Frase** | **Valor (Somatório)** | **Valor Final** |
|---|---|---|---|
| Enriquecido (escala -5 a 5) | 243: "tentar **comprar (4)** (…) site pontos" | → 4 | → 1 |
| Enriquecido (escala -5 a 5) | 810: "cliente **quer desactivar (-5)** serviços **pois** vai **sair (-3)** **sem (-1)** período **fidelização (3)**" | → -11 | → -3 |
| Enriquecido (escala -1 a 1) | 272: "cliente causa **reclama (-1)** *smartphone* tem todos dados **desligados (-1)** **mas (-1)** *assim* denota **sem (-1)** efectuar **chamadas (1)** está ser **retirado (-1)** saldo" | → -4 | → -3 |
| Enriquecido (escala -1 a 1) | 1429: "cliente **reclama (-1)** ***ainda*** **não (-1)** **contactado (1)** vai fazer **queixa (-1)** (…)" | → -3 | → -3 |

**Tabela 4.5** – Exemplos de frases classificadas pelo léxico enriquecido com sistema avançado.

Como se pode ver através da Tabela 4.5, utilizando o léxico enriquecido, mais palavas são pontuadas nas frases. Nestes exemplos, mostramos as palavras negativas a vermelho, as palavas positivas a verde, a negação a cinzento e as palavras amplificadoras a azul.

Também é possível ver que nestas frases não existem erros ortográficos nem abreviaturas, contrariamente ao que se verificava no sistema básico, pois muitas destas palavras foram corrigidas pelas regras de reescrita.

As variantes utilizadas no sistema avançado são as seguintes:

1. Léxico Afinn / Sentilex / DomainWords
2. Pré-processamento
3. Correção de palavras
4. Remoção dos textos em que apenas é incluída uma única palavra
5. Eliminação das *stopwords personalizadas*
6. Eliminação de *stopwords*
7. *Stemming*
8. Negação: Hipotese 1 com multiplicação por -1
9. Negação: Hipótese 2 com multiplicação sem efeito
10. Amplificadores
11. Atenuadores

**Figura 4.7** – Variantes do sistema avançado.

### 4.2.1 Resultados com Léxico Afinn e DomainWords

Desta forma, para ser possível visualizar o valor resultante da introdução do léxico enriquecido, *DomainWords*, relativamente aos léxicos base, apresentamos em seguida as tabelas de resultados.

| | | | | | | | | | | | | Léxico Enriquecido | | | Léxico Afinn | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Variantes** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **Taxa de acerto %** | **Custo Médio** | **Diferenças** | **Taxa de acerto %** | **Custo Médio** |
| **1** | x | | | | | | | | | | | 39,23 | 0,1523 | | 33,33 | 0,1685 |
| **2** | x | x | | | | | | | | | | 45,69 | 0,1411 | | | |
| **3** | x | x | x | | | | | | | | | 53,43 | 0,1223 | -0,0188 | | |
| **4** | x | x | x | x | | | | | | | | 53,43 | 0,1223 | | | |
| **5** | x | x | x | x | x | | | | | | | 54,43 | 0,1099 | | | |
| **6** | x | x | x | x | x | x | | | | | | 54,43 | 0,1099 | | | |
| **7** | x | x | x | x | x | x | x | | | | | 55,43 | 0,1099 | | | |
| **8** | x | x | x | x | x | x | x | x | | | | 68,54 | 0,0811 | | | |
| **9** | x | x | x | x | x | x | x | | x | | | 70,29 | 0,0712 | -0,0387 | | |
| **10** | x | x | x | x | x | x | x | | x | x | | 77,91 | 0,0587 | -0,0125 | | |
| **11** | x | x | x | x | x | x | x | | x | x | x | **80,52** | **0,0549** | | **39,33** | **0,1573** |

**Tabela 4.6** – Resultados do léxico enriquecido e Afinn com sistema avançado.

A Tabela 4.6 mostra os resultados do sistema avançado com o léxico enriquecido. Este sistema apresentou melhores resultados à medida que foram introduzidas determinadas facetas, com a exceção da variante 6 que inclui a eliminação de *stopwords* e da variante 4 que inclui a remoção dos textos com palavras únicas, em que o custo médio se manteve igual ao resultado anterior.

Os maiores contributos nos resultados devem-se à correção de palavras, com um decréscimo de 0.0188, ao tratamento da negação, com uma descida de 0.0387 e à introdução de amplificadores, com uma diminuição do custo médio de 0.0125.

Através da figura 4.8 podemos observar os resultados obtidos pelo sistema de forma mais explícita.

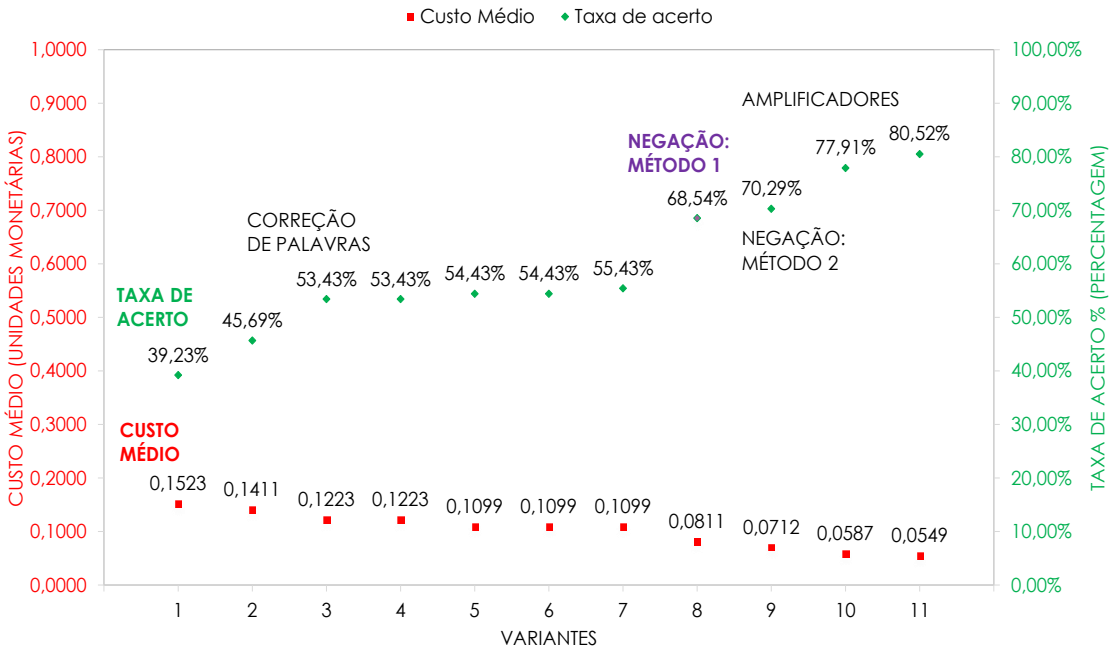



**Figura 4.8** – Gráfico de resultados do sistema avançado com léxico enriquecido.

Em suma, o léxico *DomainWords* atinge uma taxa de acerto de 80,52%, com um custo médio de 0.0549 e apresenta uma taxa de cobertura de palavras (*percentagem de palavras abrangidas pelo léxico nos documentos de texto*) na ordem dos 88,3%.

Quanto ao léxico Afinn com sistema avançado optou-se por apresentar apenas o resultado da primeira e última variantes para poder comparar com o léxico enriquecido. Em termos de resultados, atingiu uma taxa de acerto de 39,33%, com um custo médio de 0.1573, sendo a sua taxa de cobertura de palavras de 24,40 %.

### 4.2.2 Resultados com Léxico Sentilex e DomainWords

Através da Tabela 4.7, é possível observar os resultados para o léxico Sentilex com sistema avançado. Conclui-se que ao incluir pré-processamento habitual, correção de palavras, *stemming*, a hipótese 1 da negação e amplificadores, os resultados pioram.

| | | | | | | | | | | | | Léxico Sentilex | | | Léxico Enriquecido | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Variantes** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **Taxa de acerto %** | **Custo Médio** | **Diferenças** | **Taxa de acerto %** | **Custo Médio** |
| **1** | x | | | | | | | | | | | 30,70 | 0,1873 | | **28,34** | 0,2472 |
| **2** | x | x | | | | | | | | | | **31,46** | 0,1923 | | | |
| **3** | x | | x | | | | | | | | | 25,96 | 0,2022 | **0,0149** | | |
| **4** | x | | | x | | | | | | | | 30,70 | 0,1873 | | | |
| **5** | x | | | x | x | | | | | | | 30,71 | 0,1848 | **-0,0025** | | |
| **6** | x | | | x | x | x | | | | | | 30,71 | 0,1848 | | | |
| **7** | x | | | x | x | x | x | | | | | 16,23 | 0,2809 | **0,0961** | | |
| **8** | x | | | x | x | x | | x | | | | 31,34 | 0,1898 | | | |
| **9** | x | | | x | x | x | | | x | | | 31,71 | **0,1810** | **-0,0038** | | |
| **10** | x | | | x | x | x | | | x | x | | **31,46** | 0,1935 | | | |
| **11** | x | | | x | x | x | | | x | | x | 31,34 | **0,1810** | | 26,97 | **0,2247** |

**Tabela 4.7** – Resultados do léxico Sentilex e enriquecido com sistema avançado.

Notemos, que a variante 3, que inclui a correção de palavras, aumentou o custo médio em 0.0149 e a variante 7, que compreende o *stemming*, elevou o custo em 0.0961.

Pelo contrário, a variante 9, a hipótese 2 da negação, diminuiu o custo médio em 0.0038 e a variante 5, que abrange o conjunto das *stopwords personalizadas* permitiu um decréscimo do custo médio em 0.0025.

Como resultado final, tem-se uma taxa de acerto de 31,34%, com um custo médio de 0.1810, sendo a taxa de cobertura das palavras muito baixa, na ordem dos 5%.

Com o léxico enriquecido obtém-se uma taxa de acerto de 29,97% e um custo médio de 0.2247.

Esta análise mostra que estes léxicos não apresentam tão bom desempenho como os da secção anterior.

Conclui-se que o problema pode advir do facto da escala apenas tomar valores entre -1 a 1, a qual pode não ser a mais adequada, uma vez que todas as palavras assumem a mesma importância.

É possível visualizar a tabela completa destes dois léxicos no anexo D e E, bem como as matrizes de confusão e gráficos de custos no anexo F, G, H e I.

## 4.3 Léxico Otimizado para um dado domínio

A necessidade da construção de um léxico otimizado deveu-se ao facto de cada *data set* ser constituído por um novo conjunto de palavras.

Apesar de todas as palavras estarem a ser utilizadas nas experiências apresentadas anteriormente, colocou-se a hipótese do conjunto de palavras usadas num determinado conjunto de dados ser diferente quando aplicado a um outro conjunto. Além de ser uma questão de *performance* do algoritmo, também se trata de tentar otimizar um conjunto de palavras específico para cada conjunto de casos.

Esta funcionalidade é extremamente útil, pois com ela é possível eliminar palavras desnecessárias, contidas no léxico, para um dado domínio.

Portanto, foi realizado o estudo apenas para o léxico *DomainWords*, onde estão contidas 6.915 palavras, uma vez que foi o léxico que proporcionou melhor desempenho comparativamente com o Sentilex.

Também seria possível fazer o mesmo estudo para o léxico Sentilex com 82.347 expressões simples e compostas. No entanto, devido a este não apresentar tão boa *performance* como o léxico anterior e também por uma questão de tempo computacional, não foi possível efetuar essa análise.

O método utilizado para a construção destes novos léxicos apurados, nomeadamente para o conjunto de 204, 600 e 801 casos, baseou-se em seguir uma técnica que pode ser comparada à estratégia *Backward Elimination* que é usada para eliminar atributos dos classificadores.

O método consistiu em retirar uma palavra do léxico *DomainWords* e ver se o léxico com essa palavra teria ou não um custo médio superior. Dessa forma, caso fosse retirada a palavra e aumentasse o custo médio essa palavra permanecia no léxico enriquecido, caso contrário, seria removida. De seguida o processo é repetido iterativamente para todas as palavras.

Para a construção dos léxicos otimizados é utilizada a combinação de variantes número 11 encontrada anteriormente, na secção 4.2.1. Depois de obter os três novos léxicos

otimizados, pode-se afirmar que está encontrado o melhor conjunto de palavras adaptado a cada conjunto de dados.

| LÉXICO ENRIQUECIDO | Resultados Iniciais | | | Resultados Otimizados | | | Diferenças |
|---|---|---|---|---|---|---|---|
| | Taxa de acerto | Custo Médio | Total de palavras | Taxa de acerto | Custo Médio | Total de palavras | |
| 204 CASOS | 80,39% | 0,0735 | 6915 | **82,35%** | **0,0458** | 1153 | **0,0277** |
| 600 CASOS | **82,33%** | 0,0550 | 6915 | 81,50% | **0,0456** | 2733 | 0,0094 |
| 801 CASOS | 80,52% | 0,0549 | 6915 | **81,90%** | **0,0454** | 3815 | 0,0095 |

**Tabela 4.8** – Resultados do léxico enriquecido, *DomainWords*, otimizado para os diferentes *data sets*.

Como se pode observar, através da Tabela 4.8, assiste-se a uma diminuição do custo médio ao longo dos diferentes casos.

Também é possível identificar que para estes casos existe um aumento da taxa de acerto relativamente aos resultados anteriores, com a exceção dos 600 casos.

Se analisarmos o número total de palavras resultantes deste processo, observamos que estas aumentam com o acréscimo do número de casos, o que faz sentido.

Através do Anexo J, é possível analisar as respetivas matrizes de confusão e gráficos de custos para os três *data sets*.

## 4.4 Aprendizagem Automática Supervisionda

De forma, a podermos validar os modelos de classificação baseados em léxico, recorreu-se à abordagem supervisionada para que fosse possível obter uma comparação.

Na abordagem supervisionada, à partida a aprendizagem é realizada com base nos exemplos existindo intervenção humana para a construção do modelo e atribuindo desta forma classes aos exemplos. Este tipo de aprendizagem procura tendências e padrões nos conjuntos de dados.

Os algoritmos utilizados para este estudo foram: Random Forest, Neural Networks, Support Vector Machine (SVM), Decision Trees, Naive Bayes e K-Nearest Neigbours. Foram efetuadas algumas experiências com o objetivo de observar se, estes 6 algoritmos, podiam atingir melhores resultados do que a abordagem baseada em léxico.

O conjunto de documentos foi dividido em dois, ou seja, o conjunto de treino com 800 casos e um conjunto de teste com 204 casos usados para testar os classificadores e posteriormente efetuar os cálculos das medidas de desempenho, nomeadamente a taxa de acerto e o custo médio.

Na Tabela 4.9, observa-se que cada linha corresponde a uma experiência avaliada para cada algoritmo. Durante estes testes foram efetuadas diversas experiências com os vários parâmetros que os algoritmos podem assumir. Consoante essas tentativas, apenas os melhores resultados foram tidos em consideração.

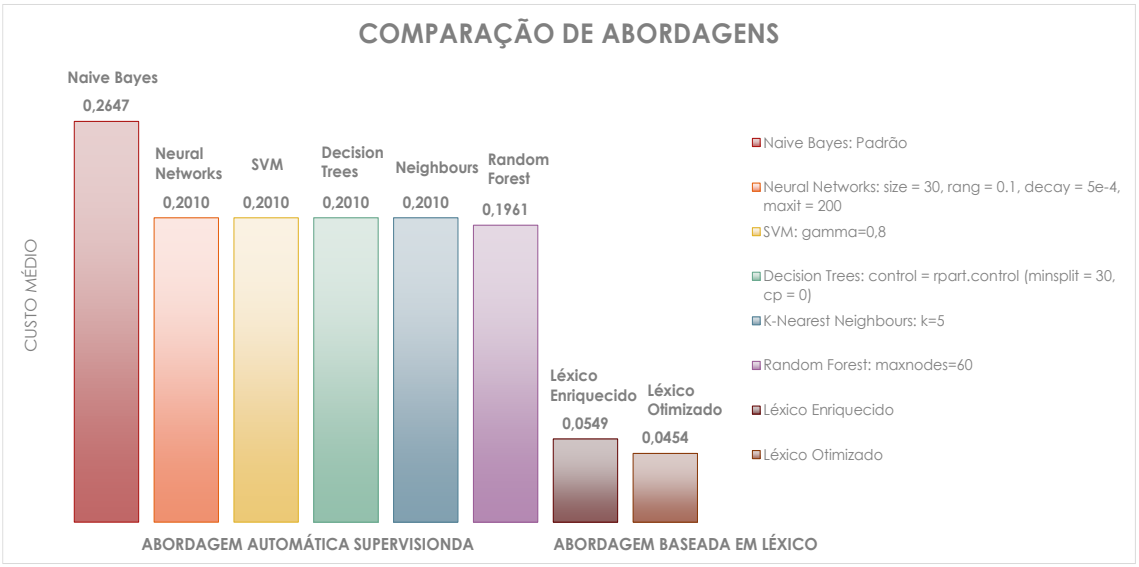



**Figura 4.9** – Resultados com aprendizagem automática supervisionada.

Constata-se que as medidas de desempenho apresentam-se muito similares para todos os algoritmos do estudo.

Observa-se que o algoritmo com melhor desempenho é o Random Forest com uma taxa de acerto de 31,37% e com custo médio de 0.1961. Pelo contrário, Naive Bayes foi o que apresentou pior desempenho comparativamente com os outros algoritmos.

Digamos que os resultados ficaram bastante aquém do que se esperava quando comparados com a abordagem baseada em léxico, em que a melhor variante teve um custo médio de 0.0458.

# CAPÍTULO 5

# CONCLUSÕES E SUGESTÕES PARA TRABALHOS FUTUROS

Neste capítulo são apresentadas as conclusões retiradas relativas a todo o trabalho efetuado nesta dissertação, bem como são apresentadas algumas considerações para que se possa melhorar e avançar com este estudo noutras vertentes.

## 5.1 Conclusões

Ao longo desta dissertação, a temática estudada inseriu-se na extração de opiniões geradas pelos utilizadores, sob a forma de comentários, escritos na língua portuguesa. Essas opiniões foram extraídas através de técnicas de extração de texto e de seguida cada documento foi avaliado para obter o valor de sentimento.

O conjunto de dados utilizado nesta dissertação é constituído por 800 documentos, nos quais os clientes contactam a empresa da área de telecomunicações, com o intuito de receberem informações ou apresentarem reclamações acerca de determinado assunto.

Para a classificação de documentos foram utilizados dois léxicos de sentimento gerais, Afinn (*inglês*) e Sentilex (*português*). Desta forma, através de algumas experiências iniciais, decidimos adotar a seguinte metodologia de trabalho, que envolve a construção de dois sistemas: básico e avançado. No sistema básico foram analisadas 4 variantes distintas, enquanto no sistema avançado foram consideradas 11. Efetuamos experiências com todas as variantes dos dois sistemas com o objetivo de identificar a melhor configuração.

Destaca-se a boa performance do léxico enriquecido, *DomainWords*, no sistema avançado, com uma taxa de acerto de 80,52% e com um custo médio de 0.0549.

Com a construção de um léxico otimizado para um dado conjunto de dados compreendeu-se que todo este processo atinge melhores resultados.

Quanto à abordagem supervisionada constatou-se que esta obteve piores resultados quando comparada com a abordagem baseada em léxico.

Poder aplicar conhecimentos e técnicas de *Text Mining* e de *Sentiment Analysis*, no domínio da Língua Portuguesa, foi sem dúvida um grande desafio. Como já foi dito, existem poucos recursos para este idioma. Por isso, foram utilizados diversas técnicas no sentido de ultrapassar esta situação, nomeadamente, através da construção de um léxico enriquecido e da mudança de pontuações de determinadas palavras, de uma lista em que as palavras são convertidas para a sua forma normalizada, de *stopwords* específicas, da negação e da lista de palavras amplificadoras e atenuadoras.

Apesar de todos estes esforços para melhorar os resultados, é desejável que os recursos disponíveis sofram melhoramentos ao longo do tempo.

## 5.2 Trabalhos Futuros

Uma linha de trabalho futuro podia envolver a construção de um corretor ortográfico automático para a língua portuguesa. Pois apesar de se poder fazer uma lista de palavras não é de todo o mais eficiente. No sentido em que se surgir uma palavra mal escrita e esta não estiver contida no dicionário, não será possível a sua identificação.

Para que esta análise seja mais sólida e esclarecedora, deveriam ser testados outros léxicos com diferentes escalas para compreender de que modo a escala afeta a atribuição do sentimento.

Mais esforço podia ser dedicado às expressões de negação, de amplificadores e de atenuadores com palavras compostas, pois só com uma melhor análise linguística é possível determinar o valor de sentimento de algumas frases. Por exemplo "*não só gosto de chocolate, como também de biscoitos*", se só estiver incluída a palavra "*não*" na negação, a primeira parte da frase supostamente é negativa e a segunda é positiva, o que na realidade não é bem assim. No exemplo apresentado, "*não só*" e "*mas também*" deveriam ser colocados como expressões amplificadoras.

Seria interessante efetuar a mesma análise abordada nesta dissertação, mas aplicada num conjunto de dados distinto.

Para uma aplicabilidade mais prática e complementar com a análise que já foi feita, poderia ser realizado um estudo para descobrir se os comentários que tiveram uma pontuação de -3 estariam diretamente relacionados com os clientes que abandonaram os serviços da empresa a curto e longo prazo.

Também se poderia estudar a possibilidade de obter o léxico enriquecido através de métodos automáticos, em vez da abordagem (*manual*) seguida nesta dissertação e efetuar uma comparação com o apresentado.

Como possível continuação desta dissertação poderia ser realizado um trabalho na área de extração de tópicos, visto que seria interessante criar um modelo que nos permitisse perceber o que é abordado em cada documento. Por exemplo, imaginemos que temos um comentário em que o cliente diz "*Quero a rescisão do contracto. Estou muito insatisfeito com os vossos serviços de internet e televisão.*". Através deste *post* seria possível extrair tópicos como "*rescisão*", "*insatisfeito*", "*internet*" e "*televisão*", em que poderíamos relacionar estes com possíveis estratégias de negócio.

# ANEXOS

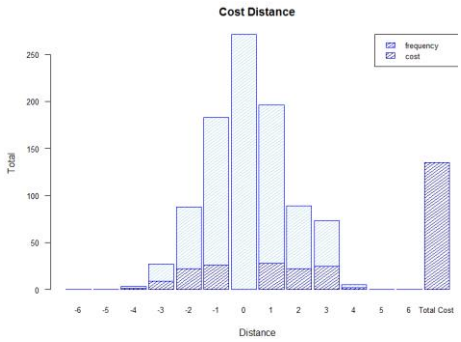
## Anexo A – Lista de *stopwords*

```
  [1] "de"          "a"           "o"           "que"         "e"
  [6] "do"          "da"          "em"          "um"          "para"
 [11] "uma"         "os"          "no"          "se"          "na"
 [16] "por"         "as"          "dos"         "como"        "ao"
 [21] "ele"         "das"         "à"           "seu"         "sua"
 [26] "nos"         "eu"          "só"          "pelo"        "pela"
 [31] "ela"         "entre"       "mesmo"       "aos"         "seus"
 [36] "quem"        "nas"         "me"          "esse"        "eles"
 [41] "você"        "essa"        "num"         "suas"        "meu"
 [46] "às"          "minha"       "numa"        "pelos"       "elas"
 [51] "qual"        "nós"         "lhe"         "deles"       "essas"
 [56] "esses"       "pelas"       "este"        "dele"        "tu"
 [61] "te"          "vocês"       "vos"         "lhes"        "meus"
 [66] "minhas"      "teu"         "tua"         "teus"        "tuas"
 [71] "nosso"       "nossa"       "nossos"      "nossas"      "dela"
 [76] "delas"       "esta"        "estes"       "estas"       "aquele"
 [81] "aquela"      "aqueles"     "aquelas"     "isto"        "aquilo"
 [86] "estamos"     "estão"       "esteve"      "estivemos"   "estiveram"
 [91] "estava"      "estávamos"   "estavam"     "estivera"    "estivéramos"
 [96] "esteja"      "estejamos"   "estejam"     "estivesse"   "estivéssemos"
[101] "estivessem"  "estiver"     "estivermos"  "estiverem"   "hei"
[106] "há"          "havemos"     "hão"         "houve"       "houvemos"
[111] "houveram"    "houvera"     "houvéramos"  "haja"        "hajamos"
[116] "hajam"       "houvesse"    "houvéssemos" "houvessem"   "houver"
[121] "houvermos"   "houverem"    "houverei"    "houverá"     "houveremos"
[126] "houverão"    "houveria"    "houveríamos" "houveriam"   "sou"
[131] "somos"       "são"         "era"         "éramos"      "eram"
[136] "fomos"       "foram"       "fôramos"     "seja"        "sejamos"
[141] "sejam"       "fosse"       "fôssemos"    "fossem"      "for"
[146] "formos"      "forem"       "serei"       "será"        "seremos"
[151] "serão"       "seria"       "seríamos"    "seriam"      "temos"
[156] "tém"         "tínhamos"    "tinham"      "tivemos"     "tiveram"
[161] "tivera"      "tivéramos"   "tenha"       "tenhamos"    "tenham"
[166] "tivesse"     "tivéssemos"  "tivessem"    "tiver"       "tivermos"
[171] "tiverem"     "terei"       "terá"        "teremos"     "terão"
[176] "teria"       "teríamos"    "teriam"
```

# Anexo B – Resultados do léxico Afinn com sistema básico

Léxico AFINN

| | sentiment | | | | | | |
|---|---|---|---|---|---|---|---|
| TrainPunctuation | -3 | -2 | -1 | 0 | 1 | 2 | 3 |
| -3 | 10 | 0 | 27 | 46 | 2 | 0 | 0 |
| -2 | 2 | 1 | 20 | 39 | 3 | 1 | 0 |
| -1 | 2 | 0 | 35 | 146 | 1 | 0 | 0 |
| 0 | 0 | 0 | 6 | 220 | 2 | 0 | 0 |
| 1 | 0 | 0 | 6 | 147 | 4 | 0 | 0 |
| 2 | 0 | 0 | 0 | 54 | 2 | 0 | 0 |
| 3 | 0 | 0 | 2 | 18 | 4 | 0 | 1 |

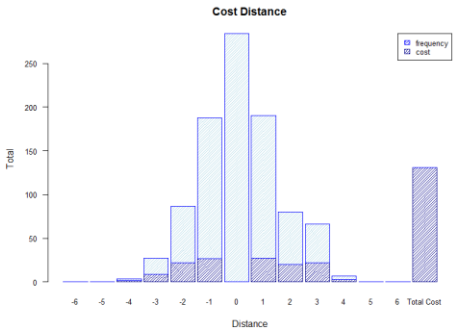



ACERTO: 33,83 %

CUSTO MÉDIO: 0.1685

Léxico AFINN + pré-processamento habitual

| | sentiment | | | | | | |
|---|---|---|---|---|---|---|---|
| TrainPunctuation | -3 | -2 | -1 | 0 | 1 | 2 | 3 |
| -3 | 15 | 0 | 26 | 41 | 3 | 0 | 0 |
| -2 | 3 | 3 | 23 | 33 | 3 | 1 | 0 |
| -1 | 2 | 0 | 44 | 137 | 1 | 0 | 0 |
| 0 | 0 | 0 | 9 | 216 | 3 | 0 | 0 |
| 1 | 0 | 0 | 8 | 144 | 5 | 0 | 0 |
| 2 | 0 | 0 | 0 | 51 | 5 | 0 | 0 |
| 3 | 0 | 0 | 2 | 18 | 4 | 0 | 1 |

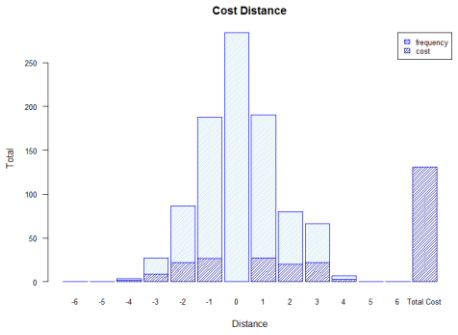



ACERTO: 35,46 %

CUSTO MÉDIO: 0.1635

Léxico AFINN + pré-processamento habitual + *stopwords*

| | sentiment | | | | | | |
|---|---|---|---|---|---|---|---|
| TrainPunctuation | -3 | -2 | -1 | 0 | 1 | 2 | 3 |
| -3 | 15 | 0 | 26 | 41 | 3 | 0 | 0 |
| -2 | 3 | 3 | 23 | 33 | 3 | 1 | 0 |
| -1 | 2 | 0 | 44 | 137 | 1 | 0 | 0 |
| 0 | 0 | 0 | 9 | 216 | 3 | 0 | 0 |
| 1 | 0 | 0 | 8 | 144 | 5 | 0 | 0 |
| 2 | 0 | 0 | 0 | 51 | 5 | 0 | 0 |
| 3 | 0 | 0 | 2 | 18 | 4 | 0 | 1 |

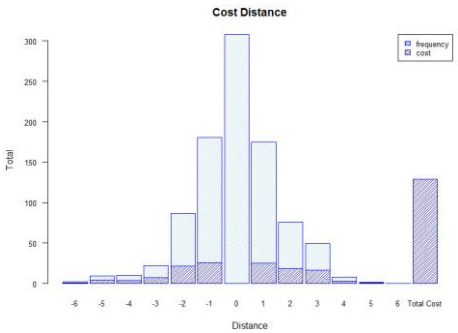



ACERTO: 35,46 %

CUSTO MÉDIO: 0.1635

Léxico AFINN + pré-processamento habitual + *stopwords* + *stemming*

| | sentiment | | | | | | |
|---|---|---|---|---|---|---|---|
| TrainPunctuation | -3 | -2 | -1 | 0 | 1 | 2 | 3 |
| -3 | 22 | 6 | 24 | 28 | 4 | 1 | 0 |
| -2 | 9 | 3 | 22 | 26 | 5 | 1 | 0 |
| -1 | 7 | 3 | 51 | 117 | 6 | 0 | 0 |
| 0 | 0 | 3 | 11 | 211 | 3 | 0 | 0 |
| 1 | 4 | 0 | 15 | 122 | 15 | 0 | 1 |
| 2 | 4 | 0 | 4 | 34 | 10 | 2 | 2 |
| 3 | 1 | 1 | 2 | 11 | 6 | 0 | 4 |



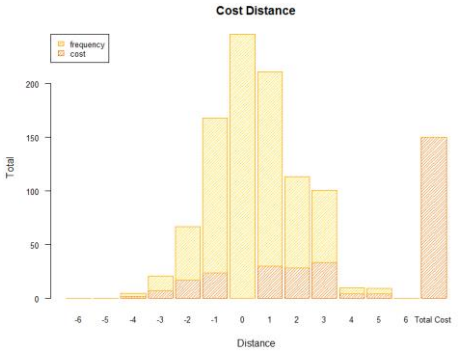

ACERTO: 38,45 %

CUSTO MÉDIO: 0.1610

# Anexo C – Resultados do léxico Sentilex com sistema básico

Léxico SENTILEX

| | sentiment | | | | | | |
|---|---|---|---|---|---|---|---|
| TrainPunctuation | -3 | -2 | -1 | 0 | 1 | 2 | 3 |
| -3 | 0 | 6 | 18 | 51 | 6 | 4 | 0 |
| -2 | 0 | 5 | 17 | 32 | 11 | 0 | 1 |
| -1 | 0 | 1 | 23 | 124 | 31 | 5 | 0 |
| 0 | 0 | 0 | 26 | 175 | 24 | 3 | 0 |
| 1 | 0 | 0 | 12 | 103 | 33 | 8 | 1 |
| 2 | 0 | 0 | 2 | 28 | 14 | 10 | 2 |
| 3 | 0 | 0 | 3 | 12 | 10 | 0 | 0 |

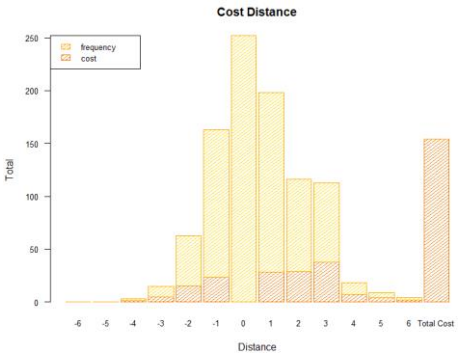



ACERTO: 30,71 %

CUSTO MÉDIO: 0.1873

Léxico SENTILEX + pré-processamento habitual

| | sentiment | | | | | | |
|---|---|---|---|---|---|---|---|
| TrainPunctuation | -3 | -2 | -1 | 0 | 1 | 2 | 3 |
| -3 | 0 | 7 | 19 | 46 | 7 | 4 | 2 |
| -2 | 0 | 7 | 16 | 26 | 13 | 3 | 1 |
| -1 | 0 | 2 | 28 | 105 | 38 | 10 | 1 |
| 0 | 0 | 0 | 27 | 169 | 24 | 2 | 6 |
| 1 | 0 | 0 | 12 | 97 | 32 | 14 | 2 |
| 2 | 0 | 0 | 2 | 26 | 13 | 11 | 4 |
| 3 | 0 | 0 | 2 | 8 | 9 | 1 | 5 |

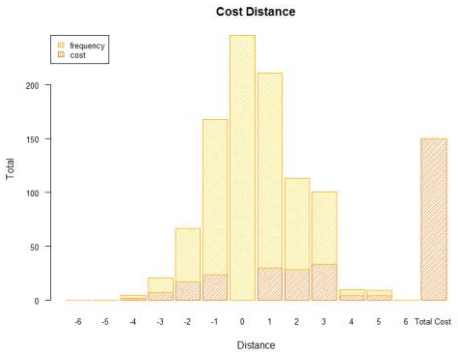



ACERTO: 31,46 %

CUSTO MÉDIO: 0.1923

Léxico SENTILEX + *stopwords*

| | sentiment | | | | | | |
|---|---|---|---|---|---|---|---|
| TrainPunctuation | -3 | -2 | -1 | 0 | 1 | 2 | 3 |
| -3 | 0 | 6 | 18 | 51 | 6 | 4 | 0 |
| -2 | 0 | 5 | 17 | 32 | 11 | 0 | 1 |
| -1 | 0 | 1 | 23 | 124 | 31 | 5 | 0 |
| 0 | 0 | 0 | 26 | 175 | 24 | 3 | 0 |
| 1 | 0 | 0 | 12 | 103 | 33 | 8 | 1 |
| 2 | 0 | 0 | 2 | 28 | 14 | 10 | 2 |
| 3 | 0 | 0 | 3 | 12 | 10 | 0 | 0 |

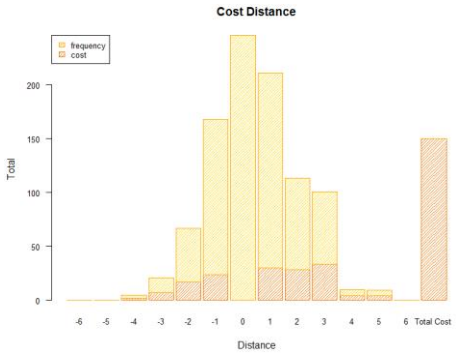



ACERTO: 30,71 %

CUSTO MÉDIO: 0.1873

Léxico SENTILEX + *stopwords* + *stemming*

| | sentiment | | | | | | |
|---|---|---|---|---|---|---|---|
| TrainPunctuation | -3 | -2 | -1 | 0 | 1 | 2 | 3 |
| -3 | 0 | 6 | 18 | 51 | 6 | 4 | 0 |
| -2 | 0 | 5 | 17 | 32 | 11 | 0 | 1 |
| -1 | 0 | 1 | 23 | 124 | 31 | 5 | 0 |
| 0 | 0 | 0 | 26 | 175 | 24 | 3 | 0 |
| 1 | 0 | 0 | 12 | 103 | 33 | 8 | 1 |
| 2 | 0 | 0 | 2 | 28 | 14 | 10 | 2 |
| 3 | 0 | 0 | 3 | 12 | 10 | 0 | 0 |




ACERTO: 30,71 %

CUSTO MÉDIO: 0.1873

# Anexo D – Resultados completos com sistema avançado: Afinn e Enriquecido

| | | | | | | | | | | | | Léxico Afinn | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Variantes** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **Taxa de acerto** | **Custo Médio** |
| **1** | x | | | | | | | | | | | 33.33% | 0.1685 |
| **2** | x | x | | | | | | | | | | 35.33% | 0.1623 |
| **3** | x | x | x | | | | | | | | | 35.46% | 0.1598 |
| **4** | x | x | x | x | | | | | | | | 35.46% | 0.1598 |
| **5** | x | x | x | x | x | | | | | | | 35.58% | **0.1573** |
| **6** | x | x | x | x | x | x | | | | | | 35.58% | **0.1573** |
| **7** | x | x | x | x | x | x | x | | | | | 39.33% | **0.1573** |
| **8** | x | x | x | x | x | x | x | x | | | | 35.33% | 0.1748 |
| **9** | x | x | x | x | x | x | x | | x | | | 36.95% | 0.1598 |
| **10** | x | x | x | x | x | x | x | | | x | | **41.07%** | 0.1586 |
| **11** | x | x | x | x | x | x | x | | | | x | 39.33% | **0.1573** |

| | | | | | | | | | | | | Léxico Enriquecido | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Variantes** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **Taxa de acerto** | **Custo Médio** |
| **1** | x | | | | | | | | | | | 39.23% | 0.1523 |
| **2** | x | x | | | | | | | | | | 45.69% | 0.1411 |
| **3** | x | x | x | | | | | | | | | 53.43% | 0.1223 |
| **4** | x | x | x | x | | | | | | | | 53.43% | 0.1223 |
| **5** | x | x | x | x | x | | | | | | | 54.43% | 0.1099 |
| **6** | x | x | x | x | x | x | | | | | | 54.43% | 0.1099 |
| **7** | x | x | x | x | x | x | x | | | | | 55.43% | 0.1099 |
| **8** | x | x | x | x | x | x | x | x | | | | 68.54% | 0.0811 |
| **9** | x | x | x | x | x | x | x | | x | | | 70.29% | 0.0712 |
| **10** | x | x | x | x | x | x | x | | x | x | | 77.91% | 0.0587 |
| **11** | x | x | x | x | x | x | x | | x | x | x | **80.52%** | **0.0549** |

| | | Classe Negativa | | | | Classe Neutra | | | | Classe Positiva | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Taxa de acerto Global | Taxa de acerto | Precisão | Recall | F1 | Taxa de acerto | Precisão | Recall | F1 | Taxa de acerto | Precisão | Recall | F1 | Custo Médio |
| **Léxico Enriquecido (escala de -5 a 5)** | 80,52% | 89,94% | 89,78% | 86,57% | 88,15% | 92,32% | 87,28% | 87,28% | 87,28% | 88,90% | 83,20% | 82,21% | 82,70% | 0,0549 |

# Anexo E – Resultados completos com sistema avançado: Sentilex e Enriquecido

| | | | | | | | | | | | | Léxico Sentilex | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Variantes** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **Taxa de acerto** | **Custo Médio** |
| **1** | x | | | | | | | | | | | 30.70% | 0.1873 |
| **2** | x | x | | | | | | | | | | 31.46% | 0.1923 |
| **3** | x | | x | | | | | | | | | 25.96% | 0.2022 |
| **4** | x | | | x | | | | | | | | 30.70% | 0.1873 |
| **5** | x | | | x | x | | | | | | | **30.71%** | 0.1848 |
| **6** | x | | | x | x | x | | | | | | **30.71%** | 0.1848 |
| **7** | x | | | x | x | x | x | | | | | 16.23% | 0.2809 |
| **8** | x | | | x | x | x | | x | | | | 31.34% | 0.1898 |
| **9** | x | | | x | x | x | | | x | | | **31.71%** | **0.1810** |
| **10** | x | | | x | x | x | | | x | x | | 31.46% | 0.1935 |
| **11** | x | | | x | x | x | | | x | | x | 31.34% | **0.1810** |

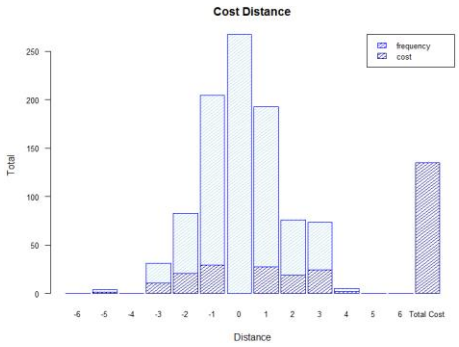
| | | Classe Negativa | | | | Classe Neutra | | | | Classe Positiva | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Taxa de acerto Global | Taxa de acerto | Precisão | Recall | F1 | Taxa de acerto | Precisão | Recall | F1 | Taxa de acerto | Precisão | Recall | F1 | Custo Médio |
| **Léxico Sentilex (escala de -1 a 1)** | 31,34% | 51,52% | 61,27% | 25,97% | 36,48% | 43,11% | 32,44% | 79,82% | 46,13% | 52,36% | 54,08% | 17,61% | 26,57% | 0,1810 |

| | | | | | | | | | | | | Léxico Enriquecido | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Variantes** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **Taxa de acerto** | **Custo Médio** |
| **1** | x | | | | | | | | | | | 28.34% | 0.2472 |
| **2** | x | x | | | | | | | | | | 22.47% | 0.2784 |
| **3** | x | | x | | | | | | | | | 24.59% | 0.2540 |
| **4** | x | | | x | | | | | | | | 28.34% | 0.2472 |
| **5** | x | | | x | x | | | | | | | **29.21%** | 0.2584 |
| **6** | x | | | x | | x | | | | | | 28.34% | 0.2472 |
| **7** | x | | | x | | x | x | | | | | 27.22% | 0.2697 |
| **8** | x | | | x | | x | | x | | | | 26.34% | 0.2397 |
| **9** | x | | | x | | x | | | x | | | 27.09% | 0.2285 |
| **10** | x | | | x | | x | | | x | x | | 28.46% | 0.2385 |
| **11** | x | | | x | | x | | | x | | x | 26.97% | **0.2247** |

# Anexo F – Resultados do léxico Afinn com sistema avançado

Léxico AFINN

| | sentiment | | | | | | |
|---|---|---|---|---|---|---|---|
| TrainPunctuation | -3 | -2 | -1 | 0 | 1 | 2 | 3 |
| -3 | 8 | 19 | 11 | 45 | 2 | 0 | 0 |
| -2 | 2 | 19 | 4 | 36 | 4 | 1 | 0 |
| -1 | 2 | 30 | 15 | 127 | 10 | 0 | 0 |
| 0 | 0 | 5 | 3 | 205 | 15 | 0 | 0 |
| 1 | 0 | 6 | 5 | 126 | 20 | 0 | 0 |
| 2 | 0 | 0 | 4 | 41 | 11 | 0 | 0 |
| 3 | 0 | 2 | 0 | 11 | 9 | 3 | 0 |

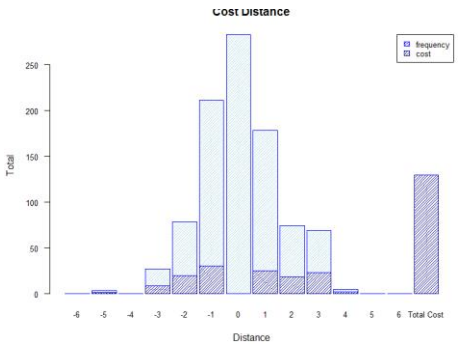



ACERTO: 33,33 %

CUSTO MÉDIO: 0.1685

Léxico AFINN + pré-processamento habitual

| | sentiment | | | | | | |
|---|---|---|---|---|---|---|---|
| TrainPunctuation | -3 | -2 | -1 | 0 | 1 | 2 | 3 |
| -3 | 13 | 19 | 11 | 39 | 3 | 0 | 0 |
| -2 | 4 | 18 | 9 | 30 | 5 | 0 | 0 |
| -1 | 2 | 34 | 24 | 109 | 13 | 2 | 0 |
| 0 | 0 | 6 | 5 | 201 | 14 | 2 | 0 |
| 1 | 0 | 7 | 8 | 116 | 24 | 2 | 0 |
| 2 | 0 | 0 | 2 | 36 | 18 | 0 | 0 |
| 3 | 0 | 2 | 0 | 9 | 7 | 4 | 3 |

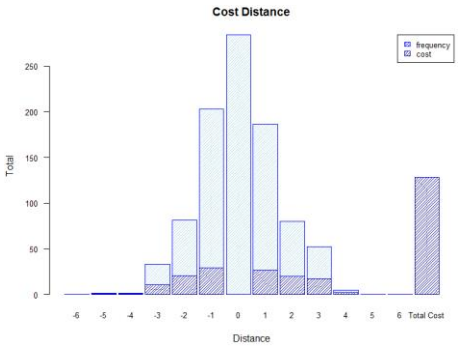



ACERTO: 35,33 %

CUSTO MÉDIO: 0.1623

Léxico AFINN + pré-processamento habitual + correção de palavras

| | sentiment | | | | | | |
|---|---|---|---|---|---|---|---|
| TrainPunctuation | -3 | -2 | -1 | 0 | 1 | 2 | 3 |
| -3 | 16 | 20 | 17 | 29 | 3 | 0 | 0 |
| -2 | 5 | 18 | 14 | 25 | 4 | 0 | 0 |
| -1 | 2 | 27 | 38 | 99 | 16 | 2 | 0 |
| 0 | 0 | 6 | 9 | 188 | 23 | 2 | 0 |
| 1 | 0 | 9 | 11 | 112 | 21 | 4 | 0 |
| 2 | 0 | 0 | 4 | 35 | 17 | 0 | 0 |
| 3 | 0 | 1 | 1 | 9 | 7 | 4 | 3 |

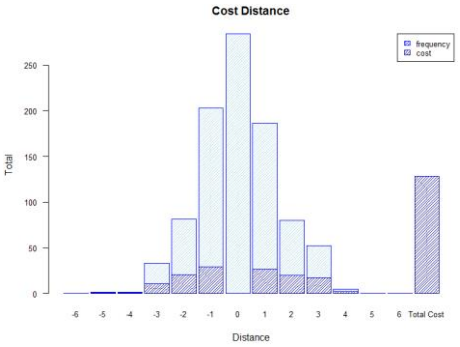



ACERTO: 35,46 %

CUSTO MÉDIO: 0.1598

Léxico AFINN + pré-processamento habitual + correção de palavras + remoção dos textos com palavras únicas

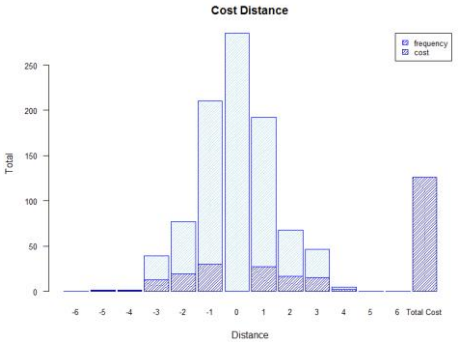
| | sentiment | | | | | | |
|---|---|---|---|---|---|---|---|
| TrainPunctuation | -3 | -2 | -1 | 0 | 1 | 2 | 3 |
| -3 | 16 | 20 | 17 | 29 | 3 | 0 | 0 |
| -2 | 5 | 18 | 14 | 25 | 4 | 0 | 0 |
| -1 | 2 | 27 | 38 | 99 | 16 | 2 | 0 |
| 0 | 0 | 6 | 9 | 188 | 23 | 2 | 0 |
| 1 | 0 | 9 | 11 | 112 | 21 | 4 | 0 |
| 2 | 0 | 0 | 4 | 35 | 17 | 0 | 0 |
| 3 | 0 | 1 | 1 | 9 | 7 | 4 | 3 |




ACERTO: 35,46 % CUSTO MÉDIO: 0.1598

Léxico AFINN + pré-processamento habitual + correção de palavras + remoção dos textos com palavras únicas + my_*stopwords*

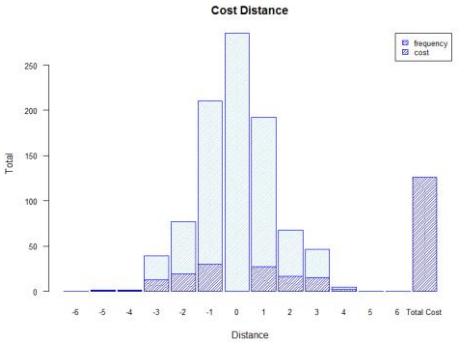
| | sentiment | | | | | | |
|---|---|---|---|---|---|---|---|
| TrainPunctuation | -3 | -2 | -1 | 0 | 1 | 2 | 3 |
| -3 | 19 | 23 | 12 | 28 | 3 | 0 | 0 |
| -2 | 4 | 20 | 13 | 26 | 3 | 0 | 0 |
| -1 | 2 | 31 | 31 | 107 | 13 | 0 | 0 |
| 0 | 0 | 6 | 5 | 197 | 20 | 0 | 0 |
| 1 | 0 | 7 | 11 | 120 | 17 | 2 | 0 |
| 2 | 0 | 0 | 4 | 35 | 17 | 0 | 0 |
| 3 | 0 | 1 | 1 | 15 | 4 | 3 | 1 |




ACERTO: 35,58 % CUSTO MÉDIO: 0.1573

Léxico AFINN + pré-processamento habitual + correção de palavras + remoção dos textos com palavras únicas + my_*stopwords* + *stopwords*

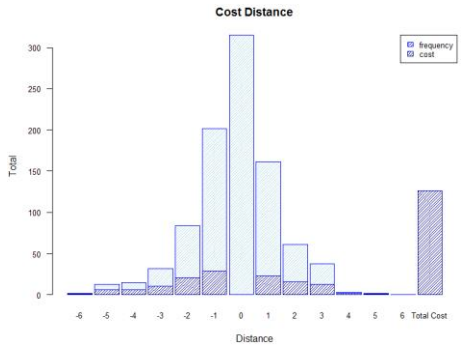
| | sentiment | | | | | | |
|---|---|---|---|---|---|---|---|
| TrainPunctuation | -3 | -2 | -1 | 0 | 1 | 2 | 3 |
| -3 | 19 | 23 | 12 | 28 | 3 | 0 | 0 |
| -2 | 4 | 20 | 13 | 26 | 3 | 0 | 0 |
| -1 | 2 | 31 | 31 | 107 | 13 | 0 | 0 |
| 0 | 0 | 6 | 5 | 197 | 20 | 0 | 0 |
| 1 | 0 | 7 | 11 | 120 | 17 | 2 | 0 |
| 2 | 0 | 0 | 4 | 35 | 17 | 0 | 0 |
| 3 | 0 | 1 | 1 | 15 | 4 | 3 | 1 |




ACERTO: 35,58 % CUSTO MÉDIO: 0.1573

Léxico AFINN + pré-processamento habitual + correção de palavras + remoção dos textos com palavras únicas + my_*stopwords* + *stopwords* + *stemming*

| | sentiment | | | | | | |
|---|---|---|---|---|---|---|---|
| TrainPunctuation | -3 | -2 | -1 | 0 | 1 | 2 | 3 |
| -3 | 30 | 20 | 11 | 21 | 2 | 1 | 0 |
| -2 | 10 | 19 | 18 | 15 | 4 | 0 | 0 |
| -1 | 7 | 34 | 47 | 78 | 18 | 0 | 0 |
| 0 | 0 | 9 | 11 | 186 | 20 | 2 | 0 |
| 1 | 6 | 8 | 19 | 97 | 25 | 2 | 0 |
| 2 | 4 | 2 | 5 | 21 | 18 | 6 | 0 |
| 3 | 1 | 3 | 1 | 8 | 7 | 3 | 2 |

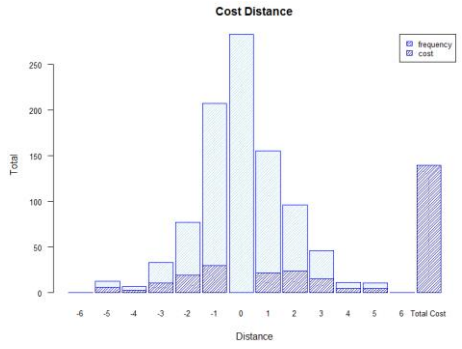



ACERTO: 39,33 %

CUSTO MÉDIO: 0.1573

Léxico AFINN + + pré-processamento habitual + correção de palavras + remoção dos textos com palavras únicas + my_*stopwords* + *stopwords* + *stemming* + negação simples

| | sentiment | | | | | | |
|---|---|---|---|---|---|---|---|
| TrainPunctuation | -3 | -2 | -1 | 0 | 1 | 2 | 3 |
| -3 | 19 | 23 | 11 | 22 | 5 | 5 | 0 |
| -2 | 3 | 21 | 18 | 17 | 4 | 2 | 1 |
| -1 | 4 | 35 | 37 | 63 | 40 | 5 | 0 |
| 0 | 0 | 8 | 16 | 176 | 24 | 4 | 0 |
| 1 | 2 | 11 | 15 | 101 | 24 | 4 | 0 |
| 2 | 3 | 1 | 3 | 24 | 20 | 4 | 1 |
| 3 | 0 | 4 | 1 | 8 | 7 | 3 | 2 |

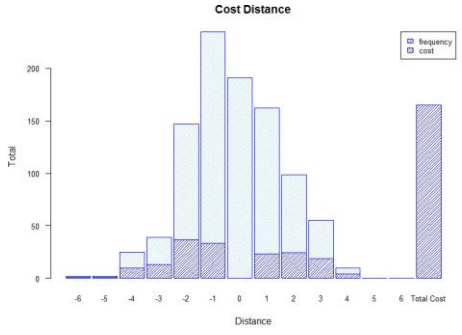



ACERTO: 35,33 %

CUSTO MÉDIO: 0.1748

Léxico AFINN + pré-processamento habitual + correção de palavras + remoção dos textos com palavras únicas + my_*stopwords* + *stopwords* + *stemming* + negação contrariada

| | sentiment | | | | | | |
|---|---|---|---|---|---|---|---|
| TrainPunctuation | -3 | -2 | -1 | 0 | 1 | 2 | 3 |
| -3 | 28 | 26 | 9 | 19 | 3 | 0 | 0 |
| -2 | 8 | 21 | 22 | 14 | 1 | 0 | 0 |
| -1 | 7 | 36 | 47 | 65 | 29 | 0 | 0 |
| 0 | 0 | 8 | 20 | 176 | 23 | 1 | 0 |
| 1 | 3 | 14 | 17 | 101 | 19 | 3 | 0 |
| 2 | 4 | 1 | 4 | 24 | 20 | 3 | 0 |
| 3 | 1 | 3 | 1 | 8 | 7 | 3 | 2 |

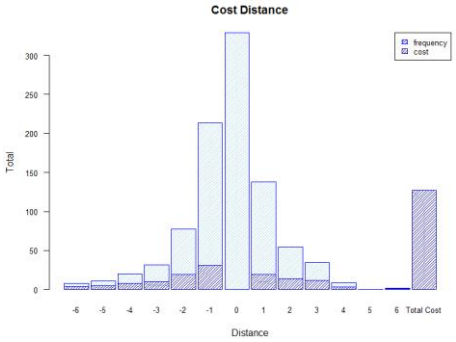



ACERTO: 36,95 %

CUSTO MÉDIO: 0.1598

Léxico AFINN + pré-processamento habitual + correção de palavras + remoção dos textos com palavras únicas + my_*stopwords* + *stopwords* + *stemming* + amplificadores

| | sentiment | | | | | | |
|---|---|---|---|---|---|---|---|
| TrainPunctuation | -3 | -2 | -1 | 0 | 1 | 2 | 3 |
| -3 | 39 | 18 | 10 | 13 | 4 | 0 | 1 |
| -2 | 22 | 17 | 10 | 10 | 6 | 1 | 0 |
| -1 | 16 | 30 | 51 | 66 | 17 | 4 | 0 |
| 0 | 2 | 7 | 13 | 183 | 19 | 4 | 0 |
| 1 | 8 | 9 | 16 | 92 | 29 | 3 | 0 |
| 2 | 5 | 4 | 3 | 13 | 22 | 7 | 2 |
| 3 | 4 | 1 | 0 | 7 | 6 | 4 | 3 |

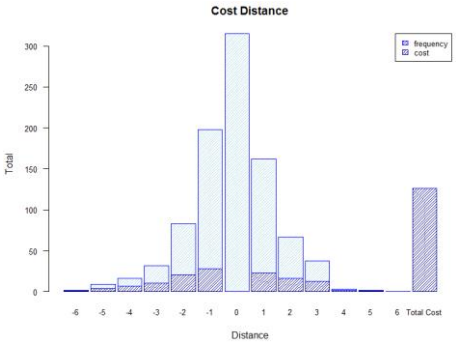



ACERTO: 41,07 %

CUSTO MÉDIO: 0.1586

Léxico AFINN + pré-processamento habitual + correção de palavras + remoção dos textos com palavras únicas + my_*stopwords* + *stopwords* + *stemming* + atenuadores

| | sentiment | | | | | | |
|---|---|---|---|---|---|---|---|
| TrainPunctuation | -3 | -2 | -1 | 0 | 1 | 2 | 3 |
| -3 | 29 | 18 | 14 | 21 | 2 | 1 | 0 |
| -2 | 9 | 19 | 19 | 15 | 4 | 0 | 0 |
| -1 | 6 | 30 | 49 | 80 | 19 | 0 | 0 |
| 0 | 0 | 9 | 12 | 185 | 20 | 2 | 0 |
| 1 | 5 | 8 | 19 | 98 | 25 | 2 | 0 |
| 2 | 3 | 3 | 5 | 21 | 18 | 6 | 0 |
| 3 | 1 | 2 | 2 | 8 | 7 | 3 | 2 |



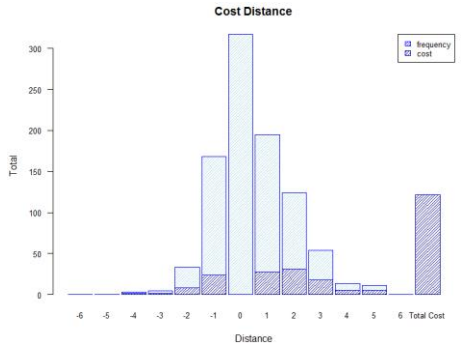

ACERTO: 39,33 %

CUSTO MÉDIO: 0.1573

# Anexo G – Resultados do léxico enriquecido (escala de -5 a 5) com sistema avançado

Léxico enriquecido

| | sentiment | | | | | | |
|---|---|---|---|---|---|---|---|
| TrainPunctuation | -3 | -2 | -1 | 0 | 1 | 2 | 3 |
| -3 | 8 | 25 | 26 | 19 | 5 | 2 | 0 |
| -2 | 2 | 13 | 17 | 22 | 6 | 2 | 4 |
| -1 | 1 | 4 | 55 | 71 | 42 | 10 | 1 |
| 0 | 0 | 2 | 17 | 165 | 40 | 3 | 1 |
| 1 | 2 | 1 | 12 | 77 | 56 | 9 | 0 |
| 2 | 0 | 0 | 0 | 7 | 31 | 13 | 5 |
| 3 | 0 | 0 | 0 | 2 | 3 | 13 | 7 |

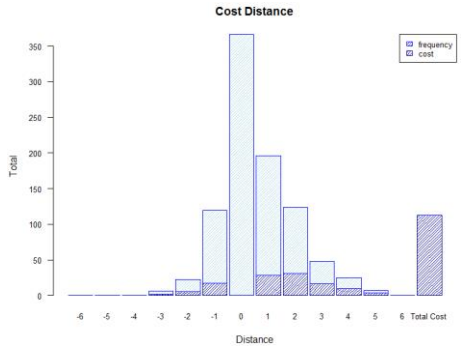



ACERTO: 39,57 %

CUSTO MÉDIO: 0.1523

Léxico enriquecido + pré-processamento habitual

| | sentiment | | | | | | |
|---|---|---|---|---|---|---|---|
| TrainPunctuation | -3 | -2 | -1 | 0 | 1 | 2 | 3 |
| -3 | 15 | 23 | 34 | 5 | 7 | 1 | 0 |
| -2 | 4 | 16 | 20 | 15 | 5 | 3 | 3 |
| -1 | 1 | 5 | 66 | 50 | 40 | 17 | 5 |
| 0 | 0 | 3 | 17 | 150 | 51 | 2 | 5 |
| 1 | 0 | 4 | 8 | 37 | 90 | 16 | 2 |
| 2 | 0 | 0 | 0 | 2 | 29 | 17 | 8 |
| 3 | 0 | 0 | 0 | 0 | 3 | 10 | 12 |

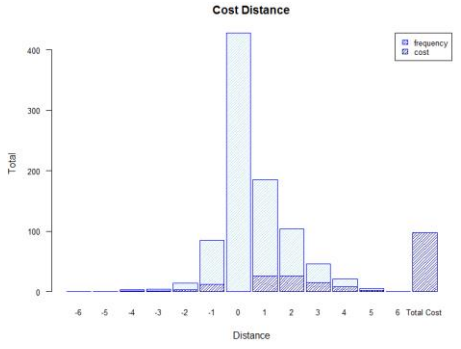



ACERTO: 45,69 %

CUSTO MÉDIO: 0.1411

Léxico enriquecido + pré-processamento habitual + correção de palavras

| | sentiment | | | | | | |
|---|---|---|---|---|---|---|---|
| TrainPunctuation | -3 | -2 | -1 | 0 | 1 | 2 | 3 |
| -3 | 26 | 33 | 19 | 4 | 2 | 1 | 0 |
| -2 | 4 | 20 | 21 | 8 | 7 | 4 | 2 |
| -1 | 0 | 9 | 70 | 42 | 41 | 15 | 7 |
| 0 | 0 | 1 | 11 | 169 | 38 | 4 | 5 |
| 1 | 2 | 3 | 9 | 21 | 100 | 16 | 6 |
| 2 | 0 | 0 | 0 | 1 | 20 | 26 | 9 |
| 3 | 0 | 0 | 0 | 0 | 0 | 8 | 17 |

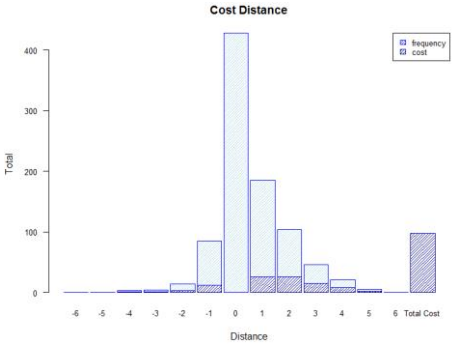



ACERTO: 53,43 %

CUSTO MÉDIO: 0.1223

Léxico enriquecido + pré-processamento habitual + correção de palavras + remoção dos textos com palavras únicas

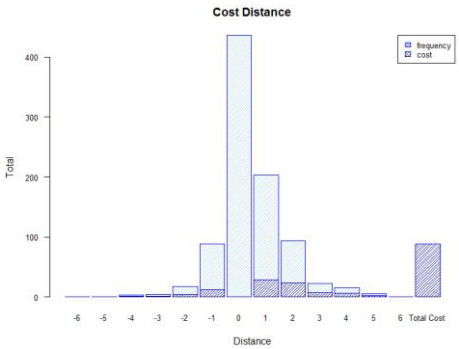
| TrainPunctuation | sentiment -3 | -2 | -1 | 0 | 1 | 2 | 3 |
|---|---|---|---|---|---|---|---|
| -3 | 26 | 33 | 19 | 4 | 2 | 1 | 0 |
| -2 | 4 | 20 | 21 | 8 | 7 | 4 | 2 |
| -1 | 0 | 9 | 70 | 42 | 41 | 15 | 7 |
| 0 | 0 | 1 | 11 | 169 | 38 | 4 | 5 |
| 1 | 2 | 3 | 9 | 21 | 100 | 16 | 6 |
| 2 | 0 | 0 | 0 | 1 | 20 | 26 | 9 |
| 3 | 0 | 0 | 0 | 0 | 0 | 8 | 17 |




ACERTO: 53,43 % CUSTO MÉDIO: 0.1223

Léxico enriquecido + pré-processamento habitual + correção de palavras + remoção dos textos com palavras únicas + my_*stopwords*

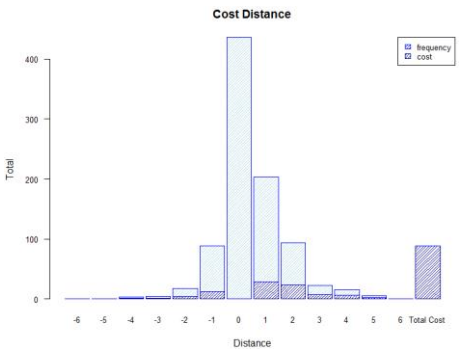
| TrainPunctuation | sentiment -3 | -2 | -1 | 0 | 1 | 2 | 3 |
|---|---|---|---|---|---|---|---|
| -3 | 30 | 30 | 21 | 1 | 2 | 1 | 0 |
| -2 | 4 | 22 | 22 | 8 | 5 | 3 | 2 |
| -1 | 0 | 5 | 77 | 53 | 37 | 8 | 4 |
| 0 | 0 | 1 | 10 | 165 | 48 | 3 | 1 |
| 1 | 2 | 3 | 8 | 26 | 104 | 13 | 1 |
| 2 | 0 | 0 | 0 | 1 | 21 | 26 | 8 |
| 3 | 0 | 0 | 0 | 0 | 3 | 10 | 12 |




ACERTO: 54,43 % CUSTO MÉDIO: 0.1099

Léxico enriquecido + pré-processamento habitual + correção de palavras + remoção dos textos com palavras únicas + my_*stopwords* + *stopwords*

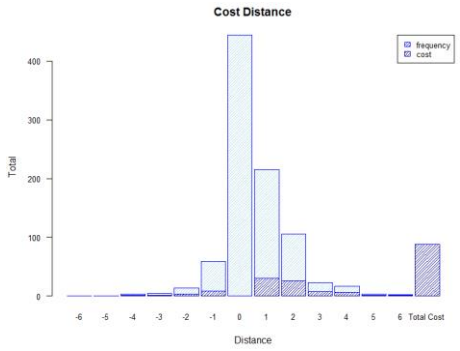
| TrainPunctuation | sentiment -3 | -2 | -1 | 0 | 1 | 2 | 3 |
|---|---|---|---|---|---|---|---|
| -3 | 30 | 30 | 21 | 1 | 2 | 1 | 0 |
| -2 | 4 | 22 | 22 | 8 | 5 | 3 | 2 |
| -1 | 0 | 5 | 77 | 53 | 37 | 8 | 4 |
| 0 | 0 | 1 | 10 | 165 | 48 | 3 | 1 |
| 1 | 2 | 3 | 8 | 26 | 104 | 13 | 1 |
| 2 | 0 | 0 | 0 | 1 | 21 | 26 | 8 |
| 3 | 0 | 0 | 0 | 0 | 3 | 10 | 12 |




ACERTO: 54,43 % CUSTO MÉDIO: 0.1099

Léxico enriquecido + pré-processamento habitual + correção de palavras + remoção dos textos com palavras únicas + my_*stopwords* + *stopwords* + *stemming*

| | sentiment | | | | | | |
|---|---|---|---|---|---|---|---|
| TrainPunctuation | -3 | -2 | -1 | 0 | 1 | 2 | 3 |
| -3 | 29 | 32 | 19 | 2 | 2 | 0 | 1 |
| -2 | 3 | 24 | 18 | 11 | 5 | 3 | 2 |
| -1 | 0 | 3 | 75 | 53 | 41 | 7 | 5 |
| 0 | 0 | 0 | 8 | 160 | 53 | 6 | 1 |
| 1 | 2 | 3 | 6 | 14 | 113 | 17 | 2 |
| 2 | 0 | 0 | 0 | 0 | 16 | 29 | 11 |
| 3 | 0 | 0 | 0 | 0 | 4 | 7 | 14 |

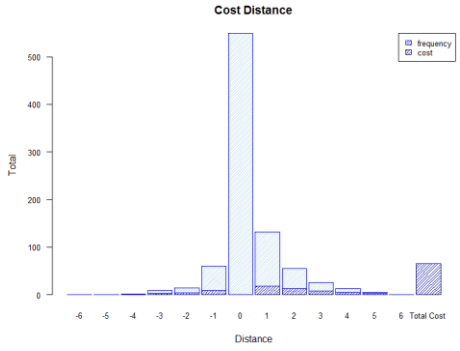



ACERTO: 55,43 %

CUSTO MÉDIO: 0.1099

Léxico enriquecido + pré-processamento habitual + correção de palavras + remoção dos textos com palavras únicas + my_*stopwords* + *stopwords* + *stemming* + negação simples

| | sentiment | | | | | | |
|---|---|---|---|---|---|---|---|
| TrainPunctuation | -3 | -2 | -1 | 0 | 1 | 2 | 3 |
| -3 | 45 | 18 | 13 | 5 | 4 | 0 | 0 |
| -2 | 1 | 28 | 22 | 7 | 4 | 1 | 3 |
| -1 | 0 | 5 | 116 | 35 | 17 | 8 | 3 |
| 0 | 2 | 0 | 8 | 197 | 20 | 1 | 0 |
| 1 | 0 | 3 | 9 | 14 | 120 | 8 | 3 |
| 2 | 0 | 0 | 0 | 0 | 15 | 31 | 10 |
| 3 | 0 | 0 | 1 | 1 | 2 | 9 | 12 |

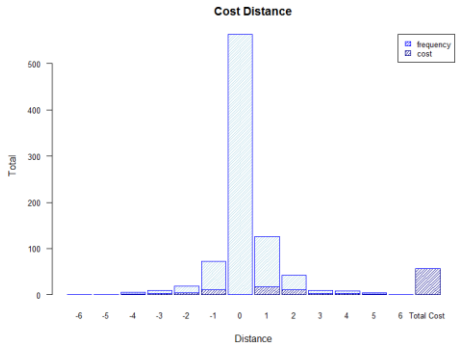



ACERTO: 68,54 %

CUSTO MÉDIO: 0.0811

Léxico enriquecido + pré-processamento habitual + correção de palavras + remoção dos textos com palavras únicas + my_*stopwords* + *stopwords* + *stemming* + negação contrariada

| | sentiment | | | | | | |
|---|---|---|---|---|---|---|---|
| TrainPunctuation | -3 | -2 | -1 | 0 | 1 | 2 | 3 |
| -3 | 47 | 23 | 12 | 1 | 2 | 0 | 0 |
| -2 | 4 | 32 | 21 | 5 | 1 | 1 | 2 |
| -1 | 0 | 7 | 123 | 34 | 14 | 4 | 2 |
| 0 | 2 | 0 | 8 | 201 | 17 | 0 | 0 |
| 1 | 2 | 3 | 11 | 15 | 119 | 6 | 1 |
| 2 | 0 | 0 | 0 | 0 | 19 | 30 | 7 |
| 3 | 0 | 0 | 1 | 1 | 3 | 9 | 11 |

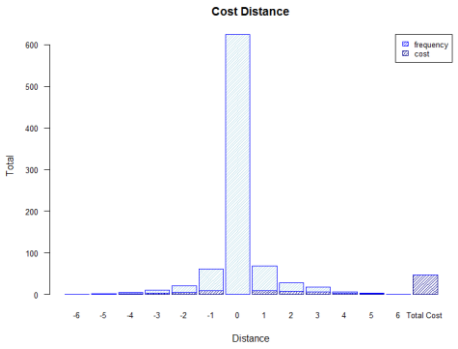



ACERTO: 70,29 %

CUSTO MÉDIO: 0.0712

Léxico enriquecido + pré-processamento habitual + correção de palavras + remoção dos textos com palavras únicas + my_*stopwords* + *stopwords* + *stemming* + negação contrariada + amplificadores

| | sentiment | | | | | | |
|---|---|---|---|---|---|---|---|
| TrainPunctuation | -3 | -2 | -1 | 0 | 1 | 2 | 3 |
| -3 | 69 | 6 | 9 | 0 | 1 | 0 | 0 |
| -2 | 7 | 55 | 0 | 0 | 1 | 1 | 2 |
| -1 | 1 | 17 | 126 | 16 | 11 | 11 | 2 |
| 0 | 2 | 1 | 11 | 194 | 20 | 0 | 0 |
| 1 | 3 | 4 | 11 | 11 | 121 | 6 | 1 |
| 2 | 0 | 0 | 0 | 0 | 3 | 42 | 11 |
| 3 | 0 | 1 | 0 | 1 | 3 | 3 | 17 |

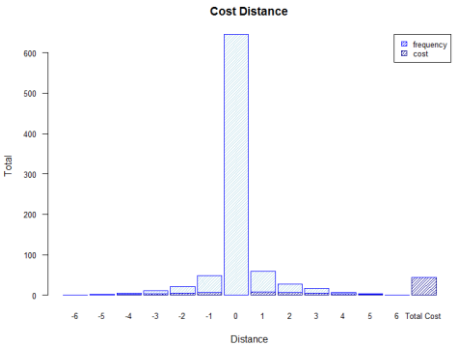



ACERTO: 77,91 %

CUSTO MÉDIO: 0.0587

Léxico enriquecido + pré-processamento habitual + correção de palavras + remoção dos textos com palavras únicas + my_*stopwords* + *stopwords* + *stemming* + negação contrariada + amplificadores + atenuadores

| | sentiment | | | | | | |
|---|---|---|---|---|---|---|---|
| TrainPunctuation | -3 | -2 | -1 | 0 | 1 | 2 | 3 |
| -3 | 70 | 4 | 10 | 0 | 1 | 0 | 0 |
| -2 | 6 | 56 | 0 | 0 | 1 | 1 | 2 |
| -1 | 1 | 8 | 135 | 18 | 10 | 10 | 2 |
| 0 | 2 | 1 | 11 | 199 | 15 | 0 | 0 |
| 1 | 3 | 4 | 11 | 10 | 125 | 3 | 1 |
| 2 | 0 | 0 | 0 | 0 | 3 | 43 | 10 |
| 3 | 0 | 1 | 0 | 1 | 3 | 3 | 17 |



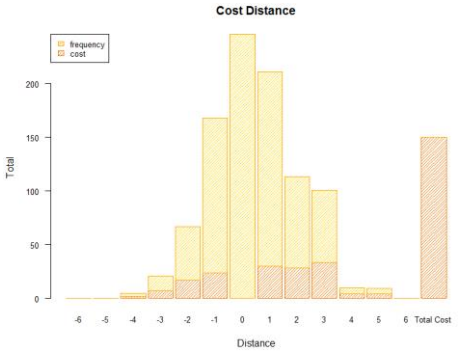

ACERTO: 80,52 %

CUSTO MÉDIO: 0.0549

# Anexo H – Resultados do léxico Sentilex com sistema avançado

Léxico SENTILEX

| | sentiment | | | | | | |
|---|---|---|---|---|---|---|---|
| TrainPunctuation | -3 | -2 | -1 | 0 | 1 | 2 | 3 |
| -3 | 0 | 6 | 18 | 51 | 6 | 4 | 0 |
| -2 | 0 | 5 | 17 | 32 | 11 | 0 | 1 |
| -1 | 0 | 1 | 23 | 124 | 31 | 5 | 0 |
| 0 | 0 | 0 | 26 | 175 | 24 | 3 | 0 |
| 1 | 0 | 0 | 12 | 103 | 33 | 8 | 1 |
| 2 | 0 | 0 | 2 | 28 | 14 | 10 | 2 |
| 3 | 0 | 0 | 3 | 12 | 10 | 0 | 0 |

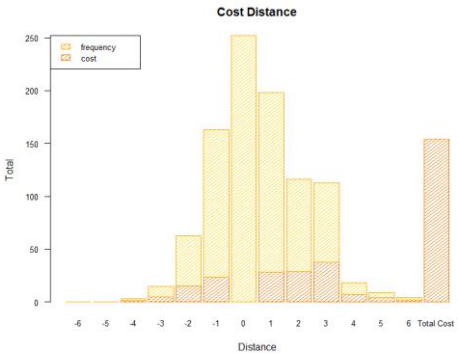



ACERTO: 30,71 %

CUSTO MÉDIO: 0.1873

Léxico SENTILEX + pré-processamento habitual

| | sentiment | | | | | | |
|---|---|---|---|---|---|---|---|
| TrainPunctuation | -3 | -2 | -1 | 0 | 1 | 2 | 3 |
| -3 | 0 | 7 | 19 | 46 | 7 | 4 | 2 |
| -2 | 0 | 7 | 16 | 26 | 13 | 3 | 1 |
| -1 | 0 | 2 | 28 | 105 | 38 | 10 | 1 |
| 0 | 0 | 0 | 27 | 169 | 24 | 2 | 6 |
| 1 | 0 | 0 | 12 | 97 | 32 | 14 | 2 |
| 2 | 0 | 0 | 2 | 26 | 13 | 11 | 4 |
| 3 | 0 | 0 | 2 | 8 | 9 | 1 | 5 |

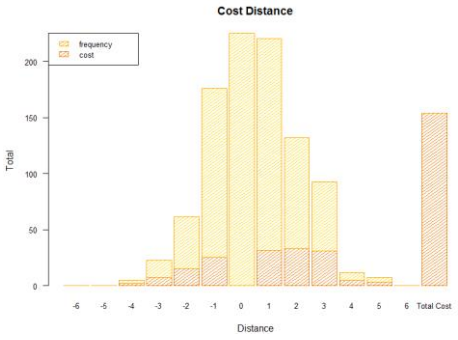



ACERTO: 31,46 %

CUSTO MÉDIO: 0.1923

Léxico SENTILEX + pré-processamento habitual + correção de palavras

| | sentiment | | | | | | |
|---|---|---|---|---|---|---|---|
| TrainPunctuation | -3 | -2 | -1 | 0 | 1 | 2 | 3 |
| -3 | 1 | 8 | 21 | 46 | 6 | 3 | 0 |
| -2 | 0 | 5 | 19 | 29 | 11 | 1 | 1 |
| -1 | 0 | 1 | 22 | 112 | 44 | 5 | 0 |
| 0 | 0 | 0 | 36 | 151 | 38 | 3 | 0 |
| 1 | 0 | 1 | 11 | 97 | 37 | 9 | 2 |
| 2 | 0 | 0 | 2 | 25 | 17 | 9 | 3 |
| 3 | 0 | 0 | 3 | 12 | 10 | 0 | 0 |

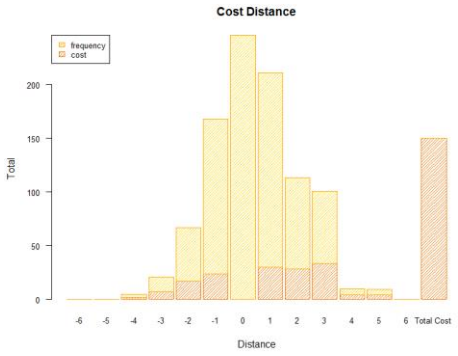



ACERTO: 28,09 %

CUSTO MÉDIO: 0.1923

Léxico SENTILEX + pré-processamento habitual + remoção dos textos com palavras únicas

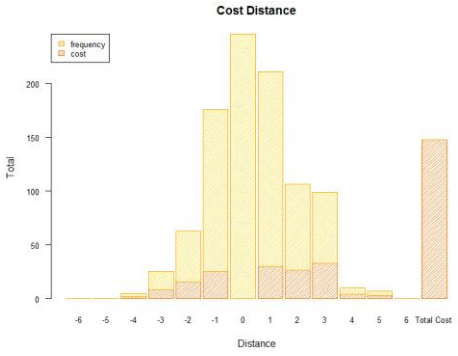
| | sentiment | | | | | | |
|---|---|---|---|---|---|---|---|
| TrainPunctuation | -3 | -2 | -1 | 0 | 1 | 2 | 3 |
| -3 | 0 | 6 | 18 | 51 | 6 | 4 | 0 |
| -2 | 0 | 5 | 17 | 32 | 11 | 0 | 1 |
| -1 | 0 | 1 | 23 | 124 | 31 | 5 | 0 |
| 0 | 0 | 0 | 26 | 175 | 24 | 3 | 0 |
| 1 | 0 | 0 | 12 | 103 | 33 | 8 | 1 |
| 2 | 0 | 0 | 2 | 28 | 14 | 10 | 2 |
| 3 | 0 | 0 | 3 | 12 | 10 | 0 | 0 |




ACERTO: 30,71 % CUSTO MÉDIO: 0.1873

Léxico SENTILEX + remoção dos textos com palavras únicas + *my_stopwords*

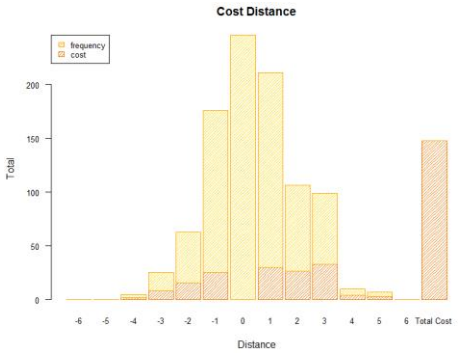
| | sentiment | | | | | | |
|---|---|---|---|---|---|---|---|
| TrainPunctuation | -3 | -2 | -1 | 0 | 1 | 2 | 3 |
| -3 | 0 | 6 | 18 | 51 | 6 | 4 | 0 |
| -2 | 0 | 4 | 19 | 32 | 11 | 0 | 0 |
| -1 | 0 | 2 | 24 | 126 | 28 | 4 | 0 |
| 0 | 0 | 0 | 26 | 180 | 20 | 2 | 0 |
| 1 | 0 | 0 | 12 | 109 | 28 | 8 | 0 |
| 2 | 0 | 0 | 2 | 28 | 14 | 10 | 2 |
| 3 | 0 | 0 | 3 | 15 | 7 | 0 | 0 |




ACERTO: 30,71 % CUSTO MÉDIO: 0.1848

Léxico SENTILEX + remoção dos textos com palavras únicas + *my_stopwords* + *stopwords*

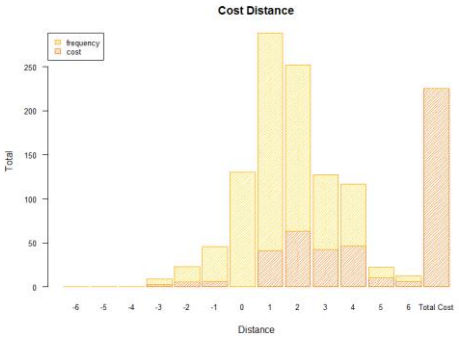
| | sentiment | | | | | | |
|---|---|---|---|---|---|---|---|
| TrainPunctuation | -3 | -2 | -1 | 0 | 1 | 2 | 3 |
| -3 | 0 | 6 | 18 | 51 | 6 | 4 | 0 |
| -2 | 0 | 4 | 19 | 32 | 11 | 0 | 0 |
| -1 | 0 | 2 | 24 | 126 | 28 | 4 | 0 |
| 0 | 0 | 0 | 26 | 180 | 20 | 2 | 0 |
| 1 | 0 | 0 | 12 | 109 | 28 | 8 | 0 |
| 2 | 0 | 0 | 2 | 28 | 14 | 10 | 2 |
| 3 | 0 | 0 | 3 | 15 | 7 | 0 | 0 |




ACERTO: 30,71 % CUSTO MÉDIO: 0.1848

Léxico SENTILEX + remoção dos textos com palavras únicas + *my_stopwords* + *stopwords* + *stemming*

| TrainPunctuation \ sentiment | -3 | -2 | -1 | 0 | 1 | 2 | 3 |
|---|---|---|---|---|---|---|---|
| -3 | 0 | 0 | 8 | 17 | 42 | 12 | 6 |
| -2 | 0 | 0 | 6 | 16 | 25 | 19 | 0 |
| -1 | 0 | 0 | 5 | 22 | 109 | 39 | 9 |
| 0 | 0 | 0 | 1 | 15 | 161 | 47 | 4 |
| 1 | 0 | 0 | 1 | 8 | 95 | 44 | 9 |
| 2 | 0 | 0 | 1 | 5 | 21 | 15 | 14 |
| 3 | 0 | 0 | 0 | 5 | 11 | 9 | 0 |

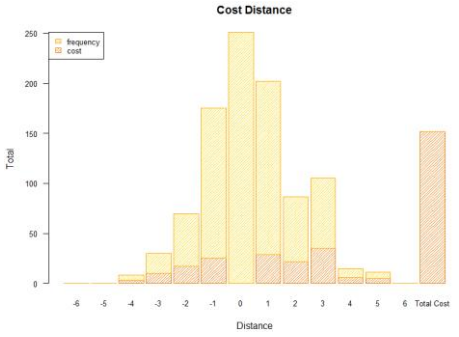



ACERTO: 16,23 %

CUSTO MÉDIO: 0.2809

Léxico SENTILEX + remoção dos textos com palavras únicas + *my_stopwords* + *stopwords* + negação simples

| TrainPunctuation \ sentiment | -3 | -2 | -1 | 0 | 1 | 2 | 3 |
|---|---|---|---|---|---|---|---|
| -3 | 1 | 3 | 15 | 52 | 8 | 6 | 0 |
| -2 | 1 | 3 | 13 | 32 | 16 | 1 | 0 |
| -1 | 0 | 1 | 34 | 129 | 18 | 2 | 0 |
| 0 | 0 | 1 | 29 | 181 | 17 | 0 | 0 |
| 1 | 0 | 0 | 15 | 108 | 25 | 9 | 0 |
| 2 | 0 | 0 | 5 | 31 | 11 | 7 | 2 |
| 3 | 0 | 0 | 5 | 15 | 5 | 0 | 0 |

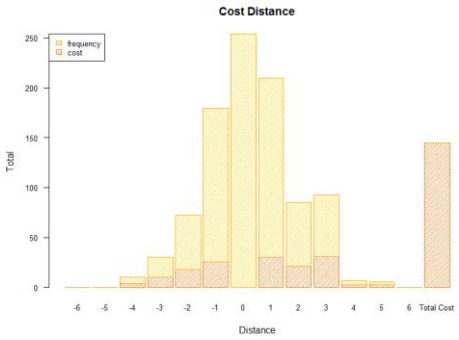



ACERTO: 31,34 %

CUSTO MÉDIO: 0.1898

Léxico SENTILEX + remoção dos textos com palavras únicas + *my_stopwords* + *stopwords* + negação contrariada

| TrainPunctuation \ sentiment | -3 | -2 | -1 | 0 | 1 | 2 | 3 |
|---|---|---|---|---|---|---|---|
| -3 | 1 | 7 | 19 | 51 | 4 | 3 | 0 |
| -2 | 2 | 6 | 19 | 30 | 9 | 0 | 0 |
| -1 | 0 | 2 | 37 | 128 | 15 | 2 | 0 |
| 0 | 0 | 1 | 30 | 181 | 16 | 0 | 0 |
| 1 | 0 | 0 | 17 | 109 | 23 | 8 | 0 |
| 2 | 0 | 0 | 5 | 32 | 11 | 6 | 2 |
| 3 | 0 | 0 | 6 | 15 | 4 | 0 | 0 |

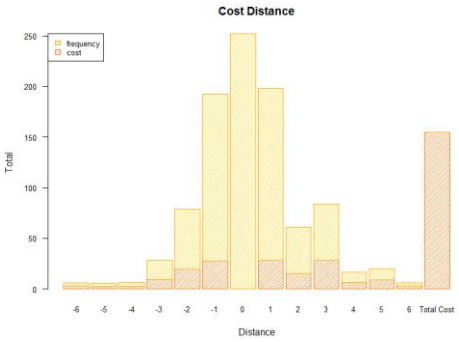



ACERTO: 31,71 %

CUSTO MÉDIO: 0.1810

Léxico enriquecido + remoção dos textos com palavras únicas + *stopwords* + negação contrariada + amplificadores

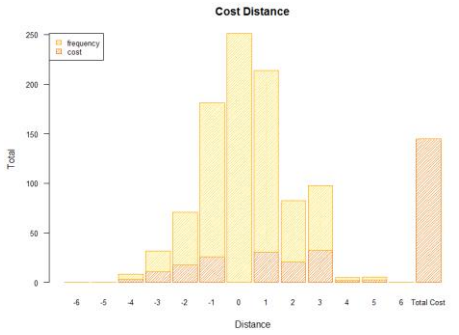
| | sentiment | | | | | | |
|---|---|---|---|---|---|---|---|
| TrainPunctuation | -3 | -2 | -1 | 0 | 1 | 2 | 3 |
| -3 | 17 | 3 | 5 | 49 | 2 | 6 | 3 |
| -2 | 19 | 0 | 8 | 29 | 4 | 1 | 5 |
| -1 | 9 | 2 | 29 | 127 | 8 | 2 | 7 |
| 0 | 2 | 0 | 29 | 181 | 15 | 0 | 1 |
| 1 | 1 | 0 | 16 | 108 | 19 | 9 | 4 |
| 2 | 2 | 0 | 3 | 31 | 7 | 5 | 8 |
| 3 | 3 | 1 | 3 | 14 | 3 | 0 | 1 |




ACERTO: 31,46 % CUSTO MÉDIO: 0.1935

Léxico enriquecido + remoção dos textos com palavras únicas + *stopwords* + negação contrariada + atenuadores

| | sentiment | | | | | | |
|---|---|---|---|---|---|---|---|
| TrainPunctuation | -3 | -2 | -1 | 0 | 1 | 2 | 3 |
| -3 | 1 | 6 | 18 | 54 | 3 | 3 | 0 |
| -2 | 2 | 5 | 18 | 32 | 9 | 0 | 0 |
| -1 | 0 | 2 | 35 | 133 | 12 | 2 | 0 |
| 0 | 0 | 1 | 29 | 182 | 16 | 0 | 0 |
| 1 | 0 | 0 | 15 | 111 | 23 | 8 | 0 |
| 2 | 0 | 0 | 5 | 33 | 11 | 5 | 2 |
| 3 | 0 | 0 | 5 | 16 | 4 | 0 | 0 |

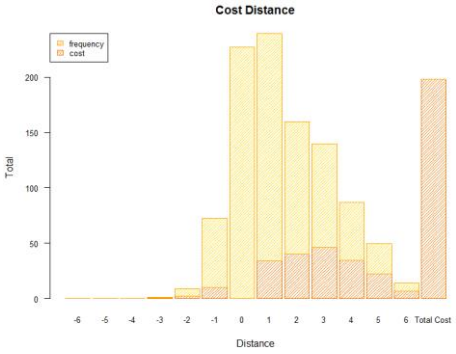



ACERTO: 31,34 % CUSTO MÉDIO: 0.1810

# Anexo I – Resultados do léxico enriquecido (escala de -1 a 1) com sistema avançado

Léxico enriquecido

| TrainPunctuation \ sentiment | -3 | -2 | -1 | 0 | 1 | 2 | 3 |
|---|---|---|---|---|---|---|---|
| -3 | 0 | 8 | 22 | 23 | 14 | 11 | 7 |
| -2 | 0 | 5 | 10 | 14 | 16 | 5 | 16 |
| -1 | 0 | 0 | 37 | 44 | 32 | 38 | 33 |
| 0 | 0 | 1 | 9 | 91 | 85 | 26 | 16 |
| 1 | 0 | 0 | 4 | 44 | 57 | 26 | 26 |
| 2 | 0 | 0 | 0 | 1 | 9 | 14 | 32 |
| 3 | 0 | 0 | 0 | 1 | 1 | 0 | 23 |

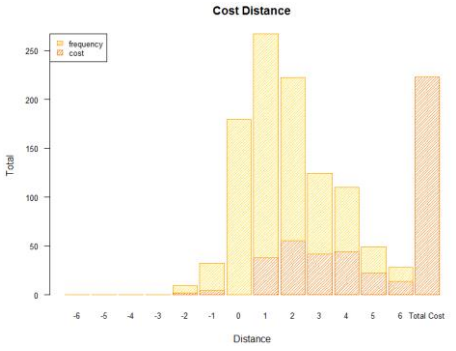



ACERTO: 28,34 %

CUSTO MÉDIO: 0.2472

Léxico enriquecido + pré-processamento habitual

| TrainPunctuation \ sentiment | -3 | -2 | -1 | 0 | 1 | 2 | 3 |
|---|---|---|---|---|---|---|---|
| -3 | 1 | 7 | 26 | 18 | 9 | 10 | 14 |
| -2 | 1 | 9 | 8 | 11 | 14 | 6 | 17 |
| -1 | 1 | 0 | 27 | 42 | 35 | 28 | 51 |
| 0 | 0 | 2 | 6 | 71 | 79 | 47 | 23 |
| 1 | 0 | 0 | 4 | 15 | 41 | 49 | 48 |
| 2 | 0 | 0 | 0 | 0 | 6 | 6 | 44 |
| 3 | 0 | 0 | 0 | 0 | 0 | 0 | 25 |

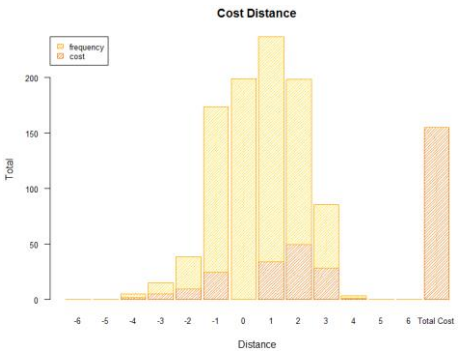



ACERTO: 22,47 %

CUSTO MÉDIO: 0.2784

Léxico enriquecido + correção de palavras

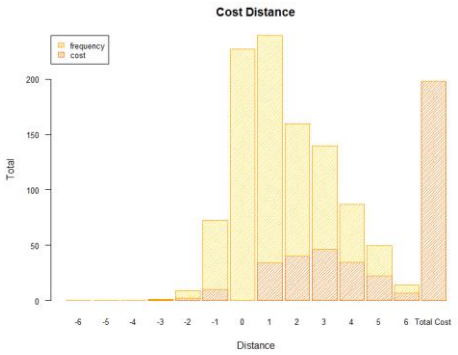
| TrainPunctuation \ sentiment | -3 | -2 | -1 | 0 | 1 | 2 | 3 |
|---|---|---|---|---|---|---|---|
| -3 | 4 | 14 | 14 | 25 | 17 | 6 | 5 |
| -2 | 0 | 6 | 10 | 16 | 8 | 9 | 17 |
| -1 | 0 | 0 | 23 | 54 | 33 | 36 | 38 |
| 0 | 0 | 1 | 5 | 88 | 88 | 19 | 27 |
| 1 | 0 | 2 | 5 | 33 | 44 | 31 | 42 |
| 2 | 0 | 0 | 0 | 0 | 5 | 10 | 41 |
| 3 | 0 | 0 | 0 | 1 | 1 | 1 | 22 |



ACERTO: 24,59 % CUSTO MÉDIO: 0.2584

Léxico enriquecido + remoção dos textos com palavras únicas

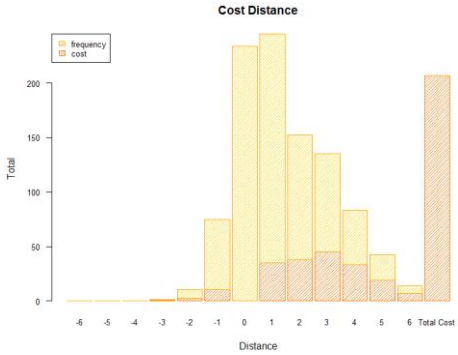
| | sentiment | | | | | | |
|---|---|---|---|---|---|---|---|
| TrainPunctuation | -3 | -2 | -1 | 0 | 1 | 2 | 3 |
| -3 | 0 | 8 | 22 | 23 | 14 | 11 | 7 |
| -2 | 0 | 5 | 10 | 14 | 16 | 5 | 16 |
| -1 | 0 | 0 | 37 | 44 | 32 | 38 | 33 |
| 0 | 0 | 1 | 9 | 91 | 85 | 26 | 16 |
| 1 | 0 | 0 | 4 | 44 | 57 | 26 | 26 |
| 2 | 0 | 0 | 0 | 1 | 9 | 14 | 32 |
| 3 | 0 | 0 | 0 | 1 | 1 | 0 | 23 |




ACERTO: 28,34 % CUSTO MÉDIO: 0.2472

Léxico enriquecido + remoção dos textos com palavras únicas + my_*stopwords*

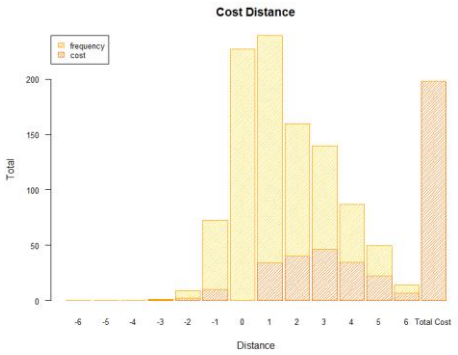
| | sentiment | | | | | | |
|---|---|---|---|---|---|---|---|
| TrainPunctuation | -3 | -2 | -1 | 0 | 1 | 2 | 3 |
| -3 | 0 | 8 | 23 | 24 | 14 | 9 | 7 |
| -2 | 0 | 5 | 10 | 15 | 17 | 5 | 14 |
| -1 | 0 | 0 | 39 | 48 | 30 | 36 | 31 |
| 0 | 0 | 1 | 9 | 93 | 88 | 24 | 13 |
| 1 | 0 | 0 | 4 | 44 | 63 | 24 | 22 |
| 2 | 0 | 0 | 0 | 1 | 9 | 14 | 32 |
| 3 | 0 | 0 | 0 | 1 | 2 | 2 | 20 |




ACERTO: 29,21 % CUSTO MÉDIO: 0.2584

Léxico enriquecido + remoção dos textos com palavras únicas + *stopwords*

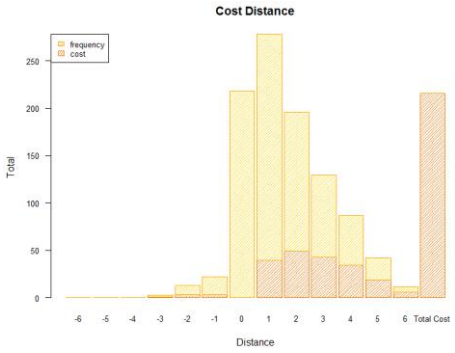
| | sentiment | | | | | | |
|---|---|---|---|---|---|---|---|
| TrainPunctuation | -3 | -2 | -1 | 0 | 1 | 2 | 3 |
| -3 | 0 | 8 | 22 | 23 | 14 | 11 | 7 |
| -2 | 0 | 5 | 10 | 14 | 16 | 5 | 16 |
| -1 | 0 | 0 | 37 | 44 | 32 | 38 | 33 |
| 0 | 0 | 1 | 9 | 91 | 85 | 26 | 16 |
| 1 | 0 | 0 | 4 | 44 | 57 | 26 | 26 |
| 2 | 0 | 0 | 0 | 1 | 9 | 14 | 32 |
| 3 | 0 | 0 | 0 | 1 | 1 | 0 | 23 |




ACERTO: 28,34 % CUSTO MÉDIO: 0.2472

Léxico enriquecido + remoção dos textos com palavras únicas + *stopwords* + *stemming*

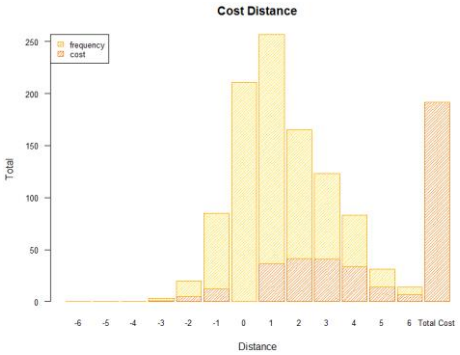
| | sentiment | | | | | | |
|---|---|---|---|---|---|---|---|
| TrainPunctuation | -3 | -2 | -1 | 0 | 1 | 2 | 3 |
| -3 | 2 | 7 | 17 | 20 | 17 | 12 | 10 |
| -2 | 1 | 3 | 8 | 12 | 12 | 8 | 22 |
| -1 | 0 | 0 | 33 | 42 | 30 | 34 | 45 |
| 0 | 0 | 1 | 7 | 86 | 80 | 33 | 21 |
| 1 | 0 | 0 | 3 | 37 | 53 | 27 | 37 |
| 2 | 0 | 0 | 0 | 0 | 5 | 12 | 39 |
| 3 | 0 | 0 | 0 | 1 | 1 | 2 | 21 |




ACERTO: 27,22 % CUSTO MÉDIO: 0.2697

Léxico enriquecido + remoção dos textos com palavras únicas + *stopwords* + negação simples

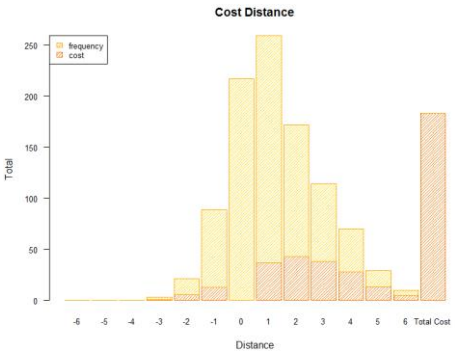
| | sentiment | | | | | | |
|---|---|---|---|---|---|---|---|
| TrainPunctuation | -3 | -2 | -1 | 0 | 1 | 2 | 3 |
| -3 | 6 | 13 | 17 | 22 | 13 | 7 | 7 |
| -2 | 2 | 4 | 7 | 14 | 16 | 13 | 10 |
| -1 | 0 | 1 | 31 | 57 | 42 | 29 | 24 |
| 0 | 0 | 3 | 12 | 90 | 81 | 27 | 15 |
| 1 | 0 | 0 | 5 | 47 | 53 | 28 | 24 |
| 2 | 0 | 0 | 1 | 3 | 9 | 9 | 34 |
| 3 | 0 | 0 | 0 | 1 | 4 | 2 | 18 |




ACERTO: 26,34 % CUSTO MÉDIO: 0.2397

Léxico enriquecido + remoção dos textos com palavras únicas + *stopwords* + negação contrariada

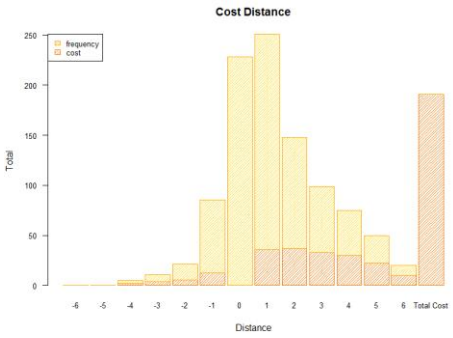
| | sentiment | | | | | | |
|---|---|---|---|---|---|---|---|
| TrainPunctuation | -3 | -2 | -1 | 0 | 1 | 2 | 3 |
| -3 | 6 | 14 | 20 | 21 | 12 | 7 | 5 |
| -2 | 2 | 4 | 9 | 19 | 15 | 8 | 9 |
| -1 | 0 | 1 | 33 | 60 | 42 | 26 | 22 |
| 0 | 0 | 3 | 12 | 93 | 81 | 25 | 14 |
| 1 | 0 | 0 | 6 | 49 | 53 | 26 | 23 |
| 2 | 0 | 0 | 1 | 3 | 10 | 10 | 32 |
| 3 | 0 | 0 | 0 | 1 | 4 | 2 | 18 |




ACERTO: 27,09 % CUSTO MÉDIO: 0.2285

Léxico enriquecido + remoção dos textos com palavras únicas + *stopwords* + negação contrariada + amplificadores

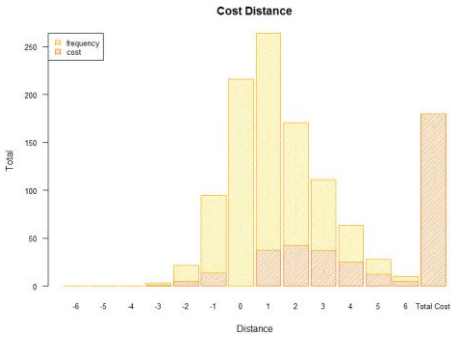
| | sentiment | | | | | | |
|---|---|---|---|---|---|---|---|
| TrainPunctuation | -3 | -2 | -1 | 0 | 1 | 2 | 3 |
| -3 | 29 | 7 | 7 | 17 | 8 | 7 | 10 |
| -2 | 6 | 7 | 9 | 12 | 8 | 4 | 20 |
| -1 | 5 | 2 | 26 | 60 | 34 | 24 | 33 |
| 0 | 4 | 2 | 10 | 93 | 80 | 22 | 17 |
| 1 | 2 | 1 | 4 | 47 | 45 | 22 | 36 |
| 2 | 0 | 1 | 1 | 2 | 7 | 8 | 37 |
| 3 | 0 | 0 | 0 | 1 | 3 | 1 | 20 |




ACERTO: 28,46 % CUSTO MÉDIO: 0.2385

Léxico enriquecido + remoção dos textos com palavras únicas + *stopwords* + negação contrariada + atenuadores

| | sentiment | | | | | | |
|---|---|---|---|---|---|---|---|
| TrainPunctuation | -3 | -2 | -1 | 0 | 1 | 2 | 3 |
| -3 | 6 | 13 | 19 | 24 | 12 | 6 | 5 |
| -2 | 2 | 3 | 8 | 24 | 13 | 7 | 9 |
| -1 | 0 | 2 | 30 | 68 | 42 | 23 | 19 |
| 0 | 0 | 3 | 12 | 97 | 81 | 21 | 14 |
| 1 | 0 | 0 | 5 | 51 | 54 | 25 | 22 |
| 2 | 0 | 0 | 1 | 3 | 11 | 10 | 31 |
| 3 | 0 | 0 | 0 | 1 | 5 | 3 | 16 |



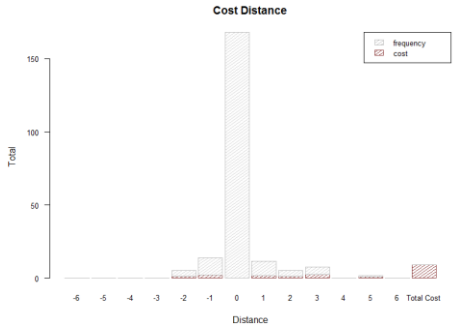

ACERTO: 26,97 % CUSTO MÉDIO: 0.2247

# Anexo J – Resultados do léxico enriquecido otimizado

Combinação n° 11: Léxico enriquecido + pré-processamento habitual + correção de palavras + remoção dos textos com palavras únicas + my_*stopwords* + *stopwords* + *stemming* + negação contrariada + amplificadores + atenuadores

[200 casos]

| | sentiment | | | | | | |
|---|---|---|---|---|---|---|---|
| TrainPunctuation | -3 | -2 | -1 | 0 | 1 | 2 | 3 |
| -3 | 24 | 1 | 4 | 1 | 0 | 1 | 0 |
| -2 | 1 | 26 | 2 | 0 | 0 | 0 | 0 |
| -1 | 0 | 2 | 28 | 2 | 0 | 4 | 0 |
| 0 | 0 | 0 | 3 | 45 | 2 | 0 | 0 |
| 1 | 0 | 0 | 4 | 4 | 33 | 1 | 0 |
| 2 | 0 | 0 | 0 | 0 | 1 | 10 | 2 |
| 3 | 0 | 0 | 0 | 0 | 0 | 1 | 2 |

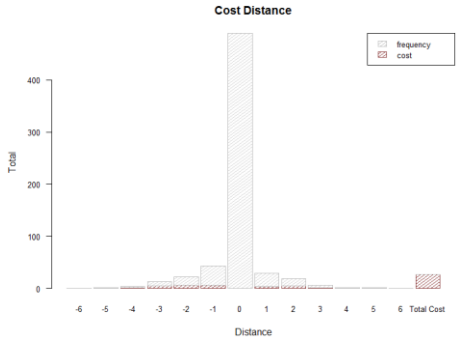



ACERTO: 82,35 %

CUSTO MÉDIO: 0.0458

[600 casos]

| | sentiment | | | | | | |
|---|---|---|---|---|---|---|---|
| TrainPunctuation | -3 | -2 | -1 | 0 | 1 | 2 | 3 |
| -3 | 55 | 3 | 4 | 1 | 0 | 0 | 0 |
| -2 | 3 | 43 | 1 | 1 | 1 | 0 | 1 |
| -1 | 1 | 7 | 107 | 14 | 8 | 2 | 1 |
| 0 | 0 | 3 | 10 | 150 | 2 | 0 | 0 |
| 1 | 1 | 6 | 11 | 8 | 92 | 1 | 1 |
| 2 | 0 | 1 | 1 | 0 | 4 | 31 | 4 |
| 3 | 0 | 1 | 0 | 2 | 2 | 5 | 11 |

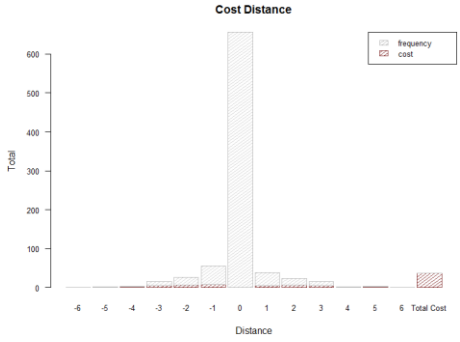



ACERTO: 81,50 %

CUSTO MÉDIO: 0.0456

[800 casos]

| | sentiment | | | | | | |
|---|---|---|---|---|---|---|---|
| TrainPunctuation | -3 | -2 | -1 | 0 | 1 | 2 | 3 |
| -3 | 71 | 3 | 10 | 1 | 0 | 0 | 0 |
| -2 | 6 | 56 | 0 | 0 | 2 | 0 | 2 |
| -1 | 2 | 8 | 143 | 16 | 7 | 7 | 1 |
| 0 | 1 | 3 | 12 | 208 | 4 | 0 | 0 |
| 1 | 1 | 7 | 12 | 11 | 122 | 3 | 1 |
| 2 | 0 | 1 | 1 | 0 | 7 | 40 | 7 |
| 3 | 0 | 1 | 0 | 1 | 3 | 4 | 16 |




ACERTO: 81,90 %

CUSTO MÉDIO: 0.454

# REFERÊNCIAS BIBLIOGRÁFICAS

Altenberg, Bengt (1998). *On the Phraseology of Spoken English: The Evidence of Recurrent Word-Combinations In A.P. Cowie (ed.), Phraseology*. Oxford: Clarendon Press.

Amine, A.; Elberrichi, Z. e Simonet, M. (2010). "Evaluation of Text Clustering Methods Using WordNet". *The International Arab Journal of Information Technology*, Vol. 7, N° 4, pp. 349–357.

Apté Chidanand; Damerau, Fred e Weiss, Sholom (1994). "Automated learning of decision rules for text categorization". *ACM Transactions on Information Systems*, Vol. 12, N° 3, pp. 233–251.

Blei, David M., Ng, Andrew Y. e Jordan, Michael I. (2003). "Latent dirichlet allocation". *The Journal of Machine Learning Research*, Vol. 3, pp. 993-1022.

Cowie, A.P. (1998). *Phraseology: Theory, Analysis, and Applications*. Oxford: Oxford University Press.

Dörre, J.; Gerstl, P. e Seiffert, R. (1999). "Text Mining: Finding Nuggets in Mountains of Textual Data". *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*. San Diego, California, USA. ACM: pp. 398-401.

Feldman R. e Sanger J. (2007). *Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press.

Feldman, R. and J. Sanger (2006). *Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*, Cambridge University Press.

Gerner, Deborah, Philip Schrodt, Ronald Francisco e Judith Weddle (1994). "The Analysis of Political Events using Machine Coded Data." International Studies Quarterly 38: pp. 91– 119.

Gomes, Helder (2012). "Text Mining: Análise de Sentimentos na classificação de notícias". Tese de Mestrado. Universidade Nova de Lisboa.

González-Ibáñez, Roberto; Muresan, Smaranda e Wacholder, Nina (2011). “Identifying sarcasm in Twitter: a closer look”. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2 (HLT '11)*. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 581-586.

Grimmer, Justin and Brandon Stewart (2013). “Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts.” Political Analysis 21: forthcoming.

Grineva, Maria; Grinev, Maxim e Lizorkin, Dmitry (2009). “Extracting key terms from noisy and multitheme documents”. *Proceedings of the 18th international conference on World wide web (WWW '09)*. ACM, New York, NY, USA, pp. 661-670.

Hofmann, Thomas (1999). “Probabilistic latent semantic indexing”. *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '99)*. ACM, New York, USA, pp. 50-57.

Honrado, A.; Leon, R.; O'Donnel, R. e Sinclair, D. (2000). "A word stemming algorithm for the Spanish language". *Proceedings of the Seventh International Symposium on String Processing Information Retrieval (SPIRE'00)*. IEEE Computer Society Washington, DC, USA, 139.

Hull, D. A. e Grefenstette, G. (1996). “A detailed analysis of english stemming algorithms”. *Technical Report TR MLTT-023*. Rank Xerox Research Centre, Meylan, France.

Jagger, Mick e Richards, Keith (1965). “(I Can't Get No) Satisfaction”. Acedido em 20 de Janeiro de 2014, em http://en.wikipedia.org/wiki/%28I_Can%27t_Get_No%29_Satisfaction.

Lin, D. e Pantel, P. (2001). “DIRT – Discovery of Inference Rules from Text”. *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining.* San Francisco, California, ACM: 323-328.

Liu, Bing (2006). *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data (Data-Centric Systems and Applications)*. Springer-Verlag New York.

Liu, Bing (2012). *Sentiment Analysis and Opinion Mining: Synthesis Lectures on Human Language Technologies*. Morgan & Claypool Publishers.

Liu, Zhiyuan; Li, Peng; Zheng, Yabin e Sun, Maosong (2009). "Clustering to find exemplar terms for keyphrase extraction". *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1 (EMNLP '09)*. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 257-266.

Manning, C. D. (2011). "Part-of-speech tagging from 97% to 100%: is it time for some linguistics?". *Proceedings of the 12th international conference on Computational linguistics and intelligent text processing - Volume Part I (CICLing'11)*, Vol. Part I. Springer-Verlag. Berlin, Heidelberg, pp. 171-189.

Mihalcea R. and Tarau, P. (2004). "TextRank: Bringing order into texts". *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pp. 404–411.

Miller, George A.; Beckwith, Richard; Fellbaum, Christiane; Gross, Derek e Miller, Katherine (1990). "WordNet: An on-line lexical database". *International Journal of Lexicography*. Vol. 3, pp. 235-244

Moral, C.; de Antonio, A.; Imbert, R. e Ramírez, J. (2014). "A survey of stemming algorithms in information retrieval". *Information Research: An International Electronic Journal*: Vol. 19, N° 1.

Nielsen, Finn (2011). "A new ANEW: Evaluation of a word list for sentiment analysis in microblogs", *Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages 718 in CEUR Workshop Proceedings*, pp. 93-98.

Nunes, L. F. et al. (2013). "Text Mining: Data Mining – INE5644". Investigação – Trabalhos em Curso, Universidade Federal de Santa Catarina.

Oakes, Michael P. (2014). *Literary Detective Work on the Computer*. John Benjamins Publishing Company.

Pang, Bo; Lee, Lillian, e Vaithyanathan, Shivakumar (2002). “Thumbs up?: sentiment classification using machine learning techniques”. *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10 (EMNLP '02)*. Association for Computational Linguistics, pp. 79-86.

Peng, Wei e Park, Dae Hoon (2011). “Generate Adjective Sentiment Dictionary for Social Media Sentiment Analysis Using Constrained Nonnegative Matrix Factorization”. *International AAAI Conference on Weblogs and Social Media*. The AAAI Press.

Polanyi, Livia e Zaenen, Annie (2006). *Computing Attitude and Affect in Text: Theory and Applications*, Vol. 20 (2006), pp. 1-10

Quinn, Kevin, Burt Monroe, Michael Crespin, Michael Colaresi e Dragomir Radev (2010). “How to Analyze Political Attention With Minimal Assumptions and Costs.” *American Journal of Political Science* 54: pp. 209–228.

Saraswat, Megha e Patel, Rahul (2014). “A Survey on Sentiment Analysis”. *International Journal of Research in Engineering Technology and Management*, Vol. 2, Issue: 6.

Sedbrook, Tod e Lightfoot, Jay M. (2010). “DEAR: A New Technique for Information Extraction and Context-Dependent Text Mining”. *Communications of the IIMA*, Vol. 10, Issue 3, Article 3, pp. 33-48

Silva, Mário J.; Carvalho, Paula e Sarmento, Luís (2012). "Building a Sentiment Lexicon for Social Judgement Mining". *Lecture Notes in Computer Science (LNCS), International Conference on Computational Processing of the Portuguese Language (PROPOR)*, Springer, pp. 218-228.

Titov, Ivan e Mcdonald, Ryan (2008). “A joint model of text and aspect ratings for sentiment summarization”. *Human Language Technologies: The 2008 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 308–316.

Tomokiyo, Takashi e Hurst, Matthew (2003). “A language model approach to keyphrase extraction”. *Proceedings of the ACL 2003 workshop on Multiword expressions: analysis,*

*acquisition and treatment - Volume 18 (MWE '03)*. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 33-40.

Turney, Peter D. e Littman, Michael L. (2003). "Measuring praise and criticism: Inference of semantic orientation from association". *ACM Transactions on Information Systems (TOIS)*, Vol. 21, Issue: 4, pp. 315-346.

Weiss, Sholom M.; Indurkhya, Nitin e Zhang, Tong (2010). *Texts in Computer Science: Fundamentals of Predictive Text Mining*. London: Springer-Verlag.

Wilson, Theresa; Wiebe, Janyce, e Hoffmann, Paul (2005). "Recognizing contextual polarity in phrase-level sentiment analysis". *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT '05)*. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 347-354.

Yang, Yiming e Pedersen, Jan O. (1997). "A Comparative Study on Feature Selection in Text Categorization". *Proceedings of the Fourteenth International Conference on Machine Learning, (ICML '97)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 412-420.

Zhu, X. e Ghahramani, Z. (2002). "Learning from Labeled and Unlabeled Data with Label Propagation". *Technical Report CMU-CALD-02-107*. Carnegie Mellon University.